Sentiment Analysis on Amharic Language-Based COVID-19 Discourse from Facebook social media comments

A Thesis presented By

Eyasu Tekle

to

The Faculty of Informatics

of

St. Mary's University

in Partial Fulfillment of the Requirements

For the Degree of Master of Science

in

Computer Science

June 2022

Acceptance

Sentiment Analysis on Amharic Language-Based COVID-19 Discourse from Facebook social media comments

By

Eyasu Tekle

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in

Computer Science

Thesis examination committee:

Alemebante Mulu (PHD)____________ ____________ ____________

Internal Examiner

____________________ ____________ ____________

External examiner

____

Dean, Faculty of Informatics

June 2022

Declaration

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other university, and all sources of material used for thesis work have been fully acknowledged.

Eyasu Tekle Lemma

Full Name of Student

Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor:

Dr. Getahun Semeon

Full Name of Advisor

Signature

Addis Ababa

Ethiopia

June 2022

ACKNOWLEDGEMENT

First of all, I would like to thank the Almighty GOD for giving me the strength, courage, and patience to accomplish this research.

Secondly, I would like to express my appreciation to my advisor Dr. Getahun Semeon for his generous help, continual advice, constructive comments, and suggestions throughout the preparation of this research paper.

Thirdly, I would like to say thanks to my friends who participated in the labeling process of this research.

Table of contents

List of Tables

List of Figures

List of Acronyms

BOW	Bag of Words
CBOW	Continuous Bag of Words
Covid-19	Coronavirus disease
FN	False Negative
FP	False Positive
MaxEnt	Maximum Entropy
ML	Machine Learning
MOH	Ministry of Health
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency Inverse Document Frequency
TN	True Negative
TP	True Positive

Abstract

The new coronavirus disease (COVID-19) outbreak from Wuhan, China in late December 2019. The virus causes respiratory infections ranging from the common cold to more serious respiratory problems. covid-19 pandemic made huge impacts on different sectors environmental, mental, economical, and industrial are some of them which the pandemic affects negatively. prior studies indicated that social media is a key tool used for gaining a huge amount of people’s opinions or sentiments towards such pandemics. Sentiment analysis is an important tool when it comes to analyzing people's expressions and thoughts on social media. The collected sentiments can be very crucial to assist public health authorities in monitoring and tracking of health information, worries, behaviors, and misinformation, and designing interventions to reduce the impact of the pandemic. In such cases, there is a need to develop a system that detects people’s opinion automatically and categorizes them as positive or negative to the guidelines given by health authorities. However, despite the importance of sentiment analysis, much investigation is not done to assess and find people’s attitudes on social media in the context of local Amharic language. The objective of this thesis is to apply sentiment analysis on Facebook social media by extracting Amharic textual comments focuses on Covid-19 and compare the performance of machine learning algorithms to find the best model. In this study, 15,000 comments regarding Covid-19 was collected and 7309 comments extracted during pre-processing stage, after which three supervised machine learning algorithms SVM, Nave Bayes, and Maximum Entropy used with feature extraction BOW, TF-IDF, and word2vec to classify sentiments expressed on comments. From which, Naïve Bayes with TF-IDF yields high results in classifying sentiments with 83.3% accuracy. The experimental evaluation shows how the proposed approach is effective.

Keywords – Sentiment analysis, Covid-19, Natural Language processing, health authorities, facebook comments

CHAPTER ONE

INTRODUCTION

1.1 Background

Coronavirus disease 2019 (COVID-19) is a virus that causes respiratory infections ranging from the common cold to more serious respiratory problems. COVID-19 was initially found in Wuhan, China, in 2019, and has since spread globally, culminating in a pandemic in 2019–2020, as designated by the World Health Organization (WHO) and the Public Health Emergency of International Concern (PHEIC) [1].

COVID-19 had made negative impacts on different sectors. We can list different vast areas in which the pandemic affects i.e. Environmental, Mental, Economical, and industrial [2], [3], [4], [5].

Social media can be effectively used in such crises to create effective strategies. The Chinese government worked efficiently using a social media platform to not only create a dedicated channel for individuals to post information on the pandemic, but also it uses the platform to mobilize citizens and nonprofit organizations to help government response and recovery efforts. [6]. By examining the comments forwarded from various citizens, the information posted on social media can be leveraged to generate beneficial insights. One of the ways that we can achieve this is by using sentiment analysis.

Sentiment analysis is a rapidly emerging field at the convergence of linguistics and computer science that aims to automatically determine the sentiment (positive or negative opinion) in the text.

Sentiment analysis can be used in different domains. One of its applications is to evaluate whether an online review (of a movie, a book, or a consumer product) is positive or negative towards the item being reviewed. Sentiment analysis is now a common tool in case of analyzing communication channels on social media carried out by companies, marketers, and political analysts [7]. Social media sentiment analysis involves applying Natural Language Processing (NLP) to analyze social media mentions from numerous sources to determine whether the user is speaking positively, negatively, or neutrally about political circumstances, religious organizations, or specific products.

Sentiment analysis in social media research can be done in a variety of ways. It can follow a rule-based, machine learning, or hybrid approach. The rule-based (lexical) approach uses a set of manually crafted positive and negative words (rules). Whereas in machine learning a sentiment analysis model learns to correctly tag a text as positive/negative using sample data. The hybrid approach works by combining rule-based and machine learning approaches to deliver more accurate results.

1.2 Motivation

The COVID-19 virus made a huge impact in our country, as evidenced by the total number of cases recorded. The statistical figure shows in Ethiopia currently, around 472,743 cases and 7,510 deaths are registered. The only appropriate method to minimize the spread of the virus, especially in developing countries like Ethiopia, is to prioritize prevention, but this is difficult to achieve. The main reason behind the increasing factor of the virus is the lack of attention and negligence in our country. One of the ways we can analyze people's opinions in this regard is by assessing the massive amount of textual information available on social media. United Nations (2020) statistics from April 8, 2020 state, that 167 countries are using national portals and social media platforms to engage people and provide vital information against covid-19 [8].

People can express their feeling on such platforms, by examining data from the social media platform Weibo, researchers found an increase in negative emotions, like anxiety, after the announcement of the disease covid-19 [9]. Other researchers discovered that nearly half of all Corona-related tweets expressed fear between mid-to end-January 2020, when the outbreak was still in its early stages [10]. During an epidemic, social media can facilitate the spread of epidemic awareness, attitudes toward control and preventive measures, emotional responses, and behaviors, as well as misinformation and rumors in the public through online interactivity [11], [12], [13].

1.3 Statement of the problem

Sentiment analysis of texts on social media can be very essential in providing useful insights by telling what is people's opinions on a certain topic. The COVID-19 pandemic is one of the hot research topics worldwide. Most government uses as an input traditional way of analyzing people's attitudes like survey-based study towards making preventive mechanisms for such viral infections.

Hagos Nigusse [14] tries to examine the covid-19 interventions and the challenges in Mekelle, Tigray, northern Ethiopia. The study uses telephone interviews, personal observation, and document reviews. Media coverage is listed as one of the major challenges as there are limited sources locally that can provide detailed information about the covid-19. Myths and misconceptions about the covid-19 also affected the covid-19 prevention efforts like eating spicy food helps to protect against the virus and consuming lemon juice, garlic and ginger to prevent the virus. The global population has been affected by widespread misinformation, which has led to an increase in despair, anxiety, and posttraumatic stress disorder [15]. The impact of a lack of information may have a greater influence on pandemic control. One study shows speculations about the virus could hamper efforts to control the spread of transmission [16].

Sentiment analysis can be very useful to get the opinion of different people on epidemics like covid-19. This can help public health authorities monitor and analyze health information, worries, habits, and disinformation, as well as plan measures to mitigate the pandemic's impact. As a result, it's critical to determine how individuals are reacting to the pandemic.

Most of the research on sentiment analysis of social media to assess people's attitudes toward covid-19 has focused on the English language and European languages.

The authors [17] analyzed tweets on covid-19 and discovered 12 topics that were divided into four primary themes: the origin, source, effects on individuals and countries, and approaches for reducing COVID-19 dissemination. The researchers suggest more work is required for sentiment analysis on social media platforms in multiple languages, as most research efforts have been devoted to English-language data.

The authors [18] created a hierarchical hybrid ensemble-based AI model for performing thematic sentiment analysis on Facebook and Twitter datasets for the covid-19 vaccine, and they finally recommend users are known to differ in their usage and preferences regarding social media platforms based on their sociodemographic profile (For example, age, socioeconomic status, and political affiliation).

So, it is important to see and examine the people's opinion on local languages like Amharic which is under-resourced language and technologically less supported language that suffers from digital corpora compared to foreign languages. There are some studies like we seen in the above and researches done on Addis Ababa city [19] but they are cross-sectional studies that are completely different from our context (survey-based study) that have limitations including small sample sizes, closed questions, and limited spatiotemporal granularity (data collected with different interval). To overcome these limitations, social media data can be used to obtain more, real-time insights into public sentiments and attitudes with considerable spatiotemporal granularity. The spatiotemporal type of analysis can be very useful to determine people's opinions in different periods. According to data from social media, researchers found positive sentiments on Facebook revealed the most notable tendency, demonstrating a continuous increasing trend from May 2020, corresponding to the start of vaccine trials and the recruitment of the first trial participant. They see a spike around mid-August 2020, which could be linked to developments in vaccine development in the UK and Russia [18].

To the best of our knowledge, there is no work done related to this regard (assess people's attitudes on covid19 using sentiment analysis on social media).

Existing models and approaches for most resource-rich languages cannot easily be adapted to Amharic due to context variations in language, culture, and technology, especially for social media communication [20]. As a result, the goal of this study is to examine diverse Amharic comments made on covid-19-related reports and news using machine learning techniques to better understand people's opinions toward the pandemic. This study will look into and answer the following research questions in order to discover a solution to the challenges mentioned above. To find a solution to the above problems, this study will investigate and answer the following research questions.

1. Which feature extraction mechanism best fit for the sentiment analysis?
2. What are the appropriate classification algorithms suitable to make effective covid-19 sentiment analysis from the data available?
3. To what extent the proposed method is efficient?

1.4 Objective

1.4.1 General Objective

- The main goal of this research is to assess and predict the individual's attitude towards covid-19 from social media using sentiment analysis.

1.4.2 Specific Objective

- To review previous related works to find suitable methods and techniques.
- To collect and prepare a dataset.
- To identify and apply different feature extraction mechanisms that are useful for the sentiment analysis.
- To develop an appropriate machine learning model for the proposed work.
- To compare and evaluate the performance of the selected technique by using quality metrics.

1.5 Methods

For achieving the objectives listed above the following methods and techniques will be applied.

1.5.1 Literature Review

To shape the research work, several kinds of literature undertaken in the field were reviewed. For this purpose, journal articles, books, and some other sources that are retrieved from the internet are assessed to understand what are the tools and techniques available to handle data and drive insights from them by making sentiment analysis, to know more about the body of knowledge, concepts, principles, and methods.

1.5.2 Research Design

In general, different research designs are applied to undertake research, they are classified into qualitative, quantitative, and design science research. The design science research technique is used in this thesis.

The production and performance of (designed) artifacts are the focus of the design science approach, with the explicit goal of improving functional performance. It can use state-of-practice applications with state-of-practice techniques and readily available components.

The Most commonly followed steps of the Design research methodology are described in the table below [21].

Table 1.1: Design Science Process

Research Step	Concerns
Identify Problem and Motivate	• Define problem • Show importance
Define Objectives of a Solution	• What would a better artifact accomplish?
Design & Development	• Artifact
Demonstration	• Find suitable context • Use artifacts to solve the problem
Evaluation	• Observe how the artifact is effective and efficient • Iterate back to design
Communication	• Scholarly publications

As shown in the table above, the first step in design research is to state the research problem and motivation. In this thesis, the research question is identified, and once the appropriate problem has been identified, a solution is proposed in the objective section (to develop a machine learning model to assess and determine people's opinions toward covid-19). Then the next stage is to build an artifact (in this step the actual machine learning model is created and the architecture is described). Finally, the artifact functionality is tested by using relevant metrics and analysis techniques.

1.5.3 Data Preparation

The major thing after identifying the problem is the data preparation step. The data were collected from social media that focuses on covid-19 related news and reports. After collecting the data then the data pre-processing steps are applied like filtering, text normalization, and removing unnecessary characters from the dataset. After that the data is analyzed, this means machine learning techniques are used to automatically learn from the data and make classification tasks after making feature extraction.

1.5.4 Tools and Techniques

In this study Machine learning approach is used to predict the attitudes of individuals behind textual information. Different ways can be followed to achieve the task supervised, unsupervised, or semi-supervised approaches can be used. The supervised machine learning approach was used for making the classification. In supervised learning, the training set (Xi, Yi) and an algorithm train a model which is capable to predict the output for every new input data. Many studies show supervised approaches can provide greater accuracy than alternative machine learning approaches. According to a survey on classification techniques for sentiment analysis and opinion mining. The use of supervised algorithms has been identified as an effective model for classifying comments with high accuracy and validity [22]. The following diagram describes the process:

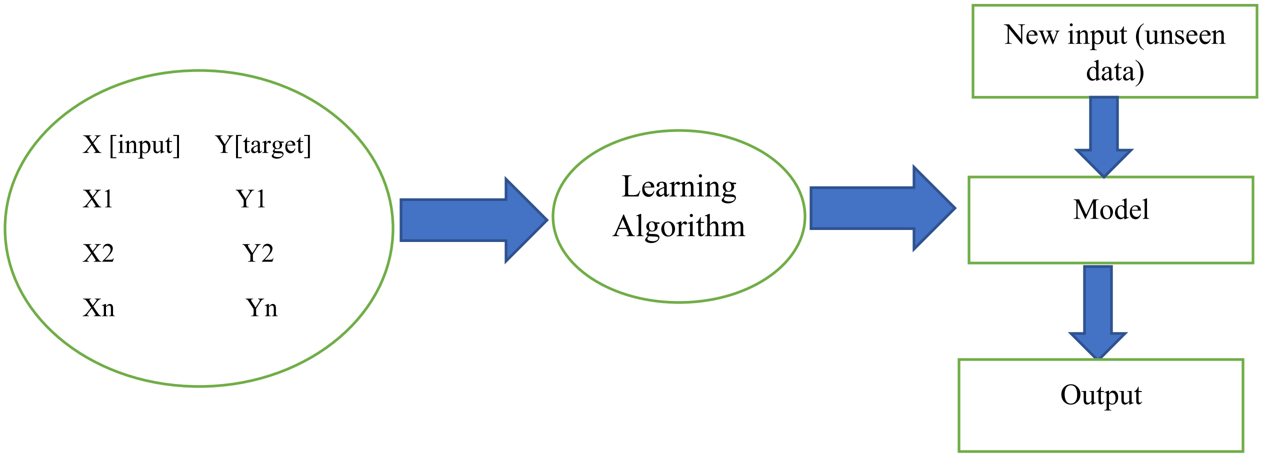


Figure 1.1: Supervised Machine Learning algorithm

1.5.5 Evaluation

In this research different machine learning based classification models were implemented and evaluated using training and testing datasets. The experimental output of the classification models was analyzed and evaluated in terms of confusion matrix, recall, precision, f-measure, and other performance metrics. As a tool, the python programming language is used to perform sentiment analysis on the dataset. Python is applied in most machine learning and deep learning areas. There are a lot of reasons python is chosen as a tool [23]:

- It has versatile libraries that can be used in machine learning and deep learning like Scikit-Learn, h2o, TensorFlow, Keras, PyTorch, etc…

- It is very high performance. Python uses high-performance libraries built in other languages that are used for parallel processing.
- It is simple and gives a lot of freedom to code due to easy syntax and readability.

1.6 Scope and Limitation

Sentiment analysis in social media has been studied from a variety of perspectives. The majority of the studies can be divided into three major categories. It can be performed at document, sentence, and entity(aspect) levels. The scope of this research is to perform sentence-level sentiment analysis to obtain sentiment expressed on comments using Amharic language, by collecting covid-19 specific comments.

- It doesn't assess emojis and idiomatic expressions.
- Limited to Amharic texts on social media.

1.7 Significance

The opinion mining process can be useful for governmental organizations to know the public perception of an event and to create an efficient prevention strategy or mitigation mechanism by determining the emotional tone of individuals.

- Monitoring and surveillance of health information, concerns, and habits can be extracted by Government organizations.
- Government organizations and agencies can design effective methods to mitigate the impact of a pandemic.

1.8 Thesis Organization

The remaining of this thesis report is organized as follows. Chapter two introduces a general overview of Amharic language, sentiment analysis, and the techniques used to perform sentiment analysis is presented, additionally, related studies performed in the area of covid-19 were reviewed. While Chapter three presents the architecture used to perform sentiment analysis and the components in the architecture are discussed in detail in the same chapter. Chapter four presents the experimental results and findings of our study. Finally, Chapter Five discusses the conclusion, recommendation, and future work.

CHAPTER TWO

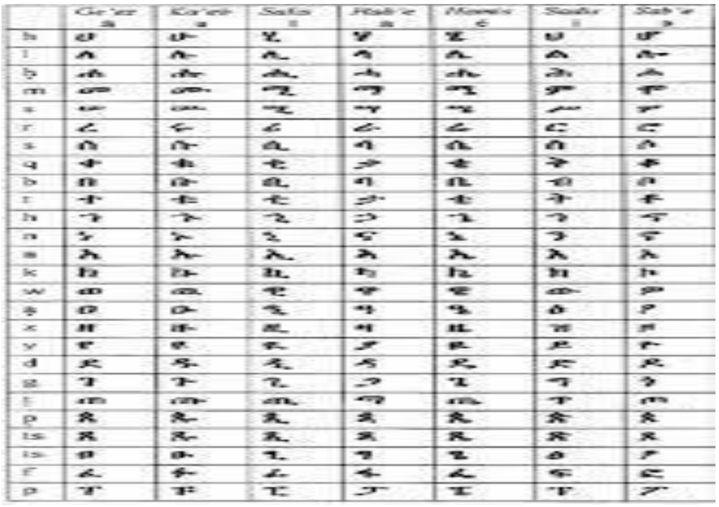
LITERATURE REVIEW

2.1 Introduction

In this chapter, first, a general overview of the Amharic language, its features, and challenges is presented, then what is sentiment analysis, its applications and methods available to perform sentiment analysis, and their efficiency are reviewed. The impact of feature extraction methods on sentiment analysis, as well as the evaluation metrics for determining the effectiveness of sentiment classification, are explored. Finally, related sentiment analysis studies performed on the covid-19 pandemic are reviewed. The purpose of this literature review is to investigate existing state of art techniques that are used to make sentiment analysis. We start by looking overview of the Amharic Language.

2.2 Amharic Language

Amharic is an Afro-Asiatic language of the southwest Semitic group. After Arabic, Amharic is the second most widely spoken Semitic language worldwide. Amharic has around 22 million native speakers. The language is used by the Ethiopian government and is also used in some parts of the country, including Addis Ababa and some areas of Southern nations, nationalities, and people (SNNP). One of the major differences between Amharic from many of the Semitic languages is the writing system.



	Ge'ez ä	Ka'eb u	Salis i	Rabi'e a	Hamis e	Sadis ɨ	Sab'e o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
ś	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
ḫ	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
ʾ	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
ʿ	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ṭ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
p	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
ts	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
ts	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Figure 2.1: Amharic Characters

Amharic language uses Fidel ("letter"), a script derived from the Ge'ez alphabet. Each consonant-vowel combination has seven forms or variations, totaling 33 fundamental characters. The language is written from left to right, unlike North Semitic languages like Arabic, Hebrew, and Syrian. [24].

The Amharic language is one of the languages, which is known as under-resourced language. Languages having one or more of the following characteristics are described as under-resourced: lack of a unique writing system or consistent orthography, linguistic expertise, and electronic resources for speech and language data [25].

When it comes to low-resource languages, sentiment analysis becomes more difficult because the basis for developing sentiment classifiers is annotated datasets, which are scarce for non-English texts [20]. Besides the lack of corpus and tools, the writing of Amharic language has also a negative impact on the performance of different machine learning algorithms. The Amharic language is one of morphologically complex language, as there are inflectional and derivational words in specific words. For instance, from the root Amharic word “ሰበር” we can derivate many words like “ሰበሩ”, “ሰበርን”,” ሰበርክ” etc.…, that make a different representation to a single word, which can increase the memory requirement of the system. So, morphological analysis is important when it comes to dealing with morphologically rich language like Amharic.

2.2.1 Unnormalized Characters

In the Amharic writing system, we can find different words having the same sounds, we can represent a group of characters by using one symbol.

Table 2.1: Same Sound Amharic Characters

Characters (Consonant)	Same sound characters
ሀ	ኀ,ሐ
ሰ	ሠ,
አ	ዓ, ዐ
ጸ	ፀ

This variation in Amharic language spellings would unnecessarily raise the number of words representing a sentence which could reduce the overall system performance and the sentiment classifier. Therefore, Amharic sentence processing should normalize word variants (spelling differences) caused by inconsistent usage of redundant characters. Like words that can be written using the above table, characters can create redundancy in our data.

For instance, we can take the word “ሰላም” and “ሠላም” which are spoken similarly yet have different orthographic interpretations. Such type of representation has to be managed to keep the performance of the classifier. The translation of noncanonical data into standard language, known as lexical normalization, has been demonstrated to improve the performance of numerous natural language processing tasks on social media. [26]

2.2.2 Compound words

Like other foreign languages, Amharic words can be combined with other words to form new meanings. This can have conventionally three forms: it can be open compounds (አየር መንገድ), close compounds (መስሪያቤት), hyphenated compounds (ስነ-ምግባር). If compound words are not managed properly, they can also make an effect on sentiment analysis. For example, if we take the word “አመለ ሸጋ” it contains a positive sentiment but it can be treated as two different words “አመለ” and “ሸጋ” that can result in a loss in sentiment value determination. So, such types of compound words should be considered to increase the performance of sentiment analysis.

2.3 Sentiment Analysis

The computational examination of people's opinions, attitudes, and feelings regarding an entity is known as sentiment analysis. Individuals, events, or subjects covered by reviews, which are textual information, can all be represented by the entity [27]. But, according to a survey on sentiment analysis and opinion mining for social multimedia, sentiment analysis may now be used for multimedia data, as photos and videos are becoming the new mediums for self-expression in social media networks [28]. For example, the author in [29] performs image sentiment analysis by using a deep learning approach and finds CNN (convolutional neural network) gives higher performance than other methods. A sentiment analysis problem is typically described as a classification problem, in which a classifier is given a text and outputs a category, indicating whether the sentiment is favorable or negative toward a topic. Because users express a range of problems online, sentiment analysis can be used in a variety of domains.

Sentiment analysis may follow three main classification levels: document, sentence, and aspect-level.

Document Level: The classifier in document-level sentiment analysis treats the entire document as a basic information unit, before classifying the opinion or sentiment as positive or negative. Document-level will be applicable if one text file is talked about a single entity (i.e., product). This kind of sentiment analysis can't detect a polarity that talks about different entities (i.e., Multiple products) existing in a document [30].

The document type of sentiment analysis has its advantages and disadvantages: The main benefit we gain from a document is an overall polarity of opinion text about a specific entity. The downside is that distinct emotions associated with different aspects of an entity cannot be retrieved independently.

Sentence Level: The classifier will deal with textual information that appeared in each sentence. This kind of analysis assumes that each sentence was made by a single person and conveys a single positive or negative viewpoint. This study is closely related to subjectivity categorization, which distinguishes between sentences that represent factual information (objective sentences) and sentences that express subjective thoughts and opinions (subjective sentences). We can state that sentence-level analysis decides first whether or not a sentence expresses an opinion then it can classify the polarity [31].

The sentence-level type of sentiment analysis has its advantages and disadvantages: subjective sentence polarity can be classified on the hand, and the disadvantage is extracting sentiment from a single sentence that contains multiple opinions can be challenging.

Aspect (Feature) Level: This level will categorize the sentiment concerning certain aspects of entities. Such type of analysis can identify the target associated with people's opinions. It starts by identifying entities and the sentiment of those entities can be extracted independently. Aspect-level sentiment analysis is concerned, not just with finding the overall sentiment associated with an entity, but also with finding the sentiment for the aspects of that entity that are discussed [32].

For Example, if we assume a restaurant sentiment analysis, the price, food, service, and other related aspects can be extracted from user comments.

Besides Sentiment analysis can give useful insights by extracting textual information mostly in social media, In contrast, it can be a challenging task. Because of the informality and special characteristics of tweets (For example., hashtags), there is a lot of noise in the data, which makes twitter analysis very different from typical sentence-level methods [33]. This is the case for other social media platforms like Facebook, WhatsApp, etc. The information gathered from these sources is filled with grammatical faults and other problems, and it cannot be processed as it is. To extract sentiments or opinions from this data, it must be preprocessed. The authors [34] explain insufficient accuracy and performance in sentiment analysis due to insufficient labeled data, dealing with a complicated sentence that requires more than sentiment words, and easy analysis are some of the obstacles that exist in performing sentiment analysis.

Sentiment analysis methods can be classified into three key categories Lexical(rule) based approach, Machine Learning, and Hybrid approach [35].

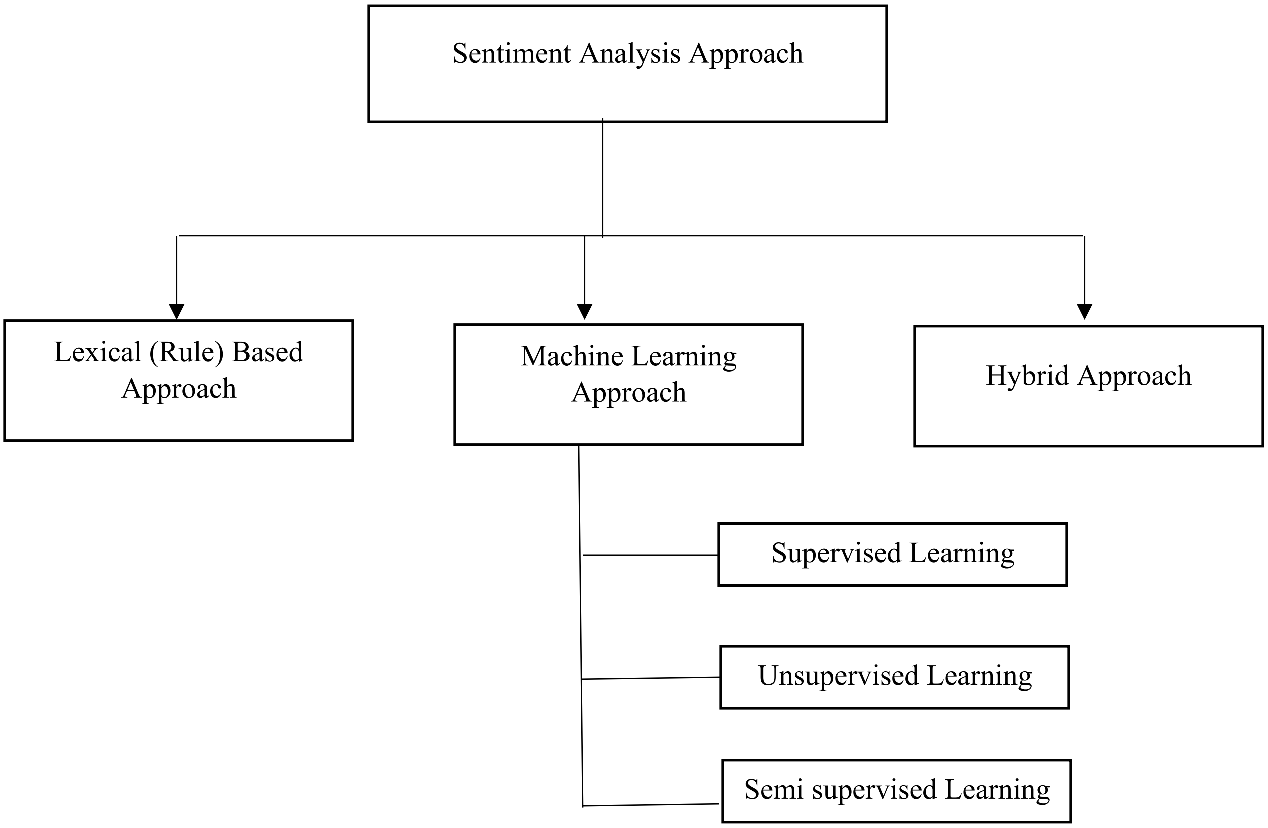

Sentiment Analysis Approach
Lexical (Rule) Based Approach
Machine Learning Approach
Hybrid Approach
Supervised Learning
Unsupervised Learning
Semi supervised Learning

Figure 2.2: Sentiment Analysis Approach

2.3.1 Rule-Based Approach

The rule-based approach also known as lexicon-based approach is a method in which we analyze the text without training or using machine learning models. The lexicon-based approach generally relies on a dictionary of opinion words, also known as a sentiment dictionary or a sentiment lexicon, to identify and determine sentiment orientation as positive or negative.

Different sentiment analysis works are done by using a rule-based approach.

The authors in [36] use VADER (Valence Aware Dictionary for Sentiment Reasoning) to analyze social media content. The goal of the study is to create a vocabulary relevant to social text in microblogs, make it available to the broader research community, and compare its effectiveness to other well-established lexicons for sentiment analysis of social media text and other domains. They build and empirically validate a gold-standard collection of lexical features (together with their related sentiment intensity measures), that are specifically attuned to sentiment in microblog-like circumstances such as emotions, acronyms, and slang. Then they analyze five general rules that incorporate grammatical and syntactical patterns for expressing and emphasizing sentiment intensity while combining these lexical aspects. Finally, VADER sentiment analysis is compared to 11 baseline models, and they discover that it generalizes better across contexts than any of their benchmarks, with F1=0.96 classification accuracy, which is even better than individual human raters F1=0.84.

The advantage of the rule-based approach is it doesn't require labeled training datasets, easy access to word lexicon and their orientation, and provides better results for the less banded domain. The limitation of this approach is we can't find the opinion words with the specific content orientation domain which is not in the lexicon and less accuracy during consideration of different domains or we can say unable to generalize to other domains [22].

2.3.2 Machine Learning Approach

The machine learning-based sentiment analysis depends on existing datasets then the model can learn how to perform new tasks without being expressly programmed to perform them. When there is a finite collection of positive/negative classifications, such as in sentiment analysis, machine learning can employ a supervised technique. To train classifiers, this technique requires labeled data [37]. The automatic classifier learns the different properties of sentiments appearing on comments using a training set, and the test set is used to validate the sentiment classifier. When finding labeled training documents is challenging, an unsupervised machine learning strategy can be employed instead.

An unsupervised machine learning approach on the other hand can be used when it is difficult to find labeled training documents. This method doesn't consist of a category and is therefore used to perform clustering tasks. The main problem of this approach is the requirement of the big magnitude of the training data. Another issue with this strategy is the development of the disjointed subject because of the non-correlation of the model with the human judgments [34].

There are machine learning algorithms that are mostly used in the area of sentiment analysis. We try to review some commonly used algorithms.

2.3.2.1 Naïve Bayes

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks including sentiment analysis. The work in [38] used the Naïve Bayes algorithm to automatically identify the contextual polarity for a large subset of sentiment expressions. Movie reviews and hotel reviews are used as a dataset 5000 positive and negative reviews are from each domain. The Naïve Bayes approach is compared with the k-NN (k-Nearest Neighbour) classifier and it outperforms by giving 80% accuracy for movie reviews.

In naïve Bayes, the classifier assumes each feature is independent of others this means the value of one feature won't affect the other feature. Not only the features, but the samples are also assumed to be independent with each other which means the probability of one observation does not affect the probability of another observation [39]. The equation of the Bayes theorem is

$\mathrm{P(X|Y)} = \frac{P(Y|X)P(X)}{P(Y)}$, where P(X|Y) is the probability of occurring X if it is known Y. P(Y|X) is the chance of occurring Y if it is known X. P(X) is the probability of occurring X and P(Y) is the probability of occurring Y.

Advantage

- It works quickly and saves a lot of time
- It requires less training data to perform well as long as the assumptions are satisfied
- It is more suitable for categorical data than numerical.

Disadvantage

- Strong violations of the independence assumptions, as well as non-linear classification issues, can cause naive Bayes classifiers to perform poorly.

2.3.2.2 Maximum Entropy

Max Entropy is a probabilistic classifier like Naïve Bayes that belongs to the class exponential models. But unlike the Naïve Bayes classifier, the model doesn't assume the features are conditionally independent of each other. The MaxEnt is based on the Principle of Maximum Entropy, and it chooses the model with the highest entropy from all the models that fit our training data. The Max Entropy classifier is also utilized in Sentiment Analysis and other text classification applications. The authors in [37] used different machine learning algorithms including the Maximum Entropy method to analyze Twitter data. The Twitter dataset used in the research is already labeled, and to improve the quality of the model preprocessing is done on the dataset. Then feature extraction is applied before the classification task to extract the aspect(adjectives) from the data. 19340 total dataset is used and 18340 of them is used in training and the remaining for testing. The maximum entropy method gives 83.8% accuracy.

Advantage

- It can be used when we don't have any know-how about the prior distributions or when we can't assume the conditional independence of the features

Disadvantage

- It requires more time to train the model compared to Naïve Bayes.

2.3.2.3 Support Vector Machine

Support Vector Machine (SVM) is the most popular and commonly used supervised machine learning algorithm. The goal of a support vector machine is to locate a hyperplane in an N-dimensional space, where N denotes the number of features that distinguish data points. SVM has a wider range of applications in the area of NLP. i.e. Opinion Mining, Information Extraction, Part of Speech tagging [[40], [41], [42]] are some areas in which SVM is used. The work in [43] attempts to evaluate the performance of SVM for textual data polarity detection. The WEKA tool is used for performance evaluation and classification. For tweets about self-driving cars and Apple products, two pre-labeled data sets are used. Positive and negative tags are pre-labeled in each dataset. After normalization is done on the data SVM is used for classification. The result shows that SVM gives an accuracy of 59.91% and 71.2% on the self-driving cars and apple products dataset respectively.

The SVM algorithm first gets the input data and tries to create a line or hyperplane that separates our target class as much as possible. In the case of Sentiment analysis SVM tries to create a line to separate the positive and negative classes. The SVM learns how essential each of the training data points is in representing the decision border between the two classes during training. Only a subset of the training points is typically used to define the decision boundary: those that are on the class boundary are called support vectors, which gives the support vector machine its name [44].

Advantage

- It works well with a clear margin of separation.
- SVM Classifiers offer good accuracy and perform faster prediction compared to the Naïve Bayes algorithm.
- They also require less memory in the decision phase because they only use a subset of training points.

Disadvantage

- When we have a large data collection, it doesn't perform well because the needed training time is longer.
- When the data set target classes overlap, it doesn't function well.

2.3.2.4 Neural Networks

Neural Networks are a subset of Machine Learning algorithms and the base for deep learning algorithms. The model is built by trying to simulate the way how biological neurons pass signals to one another in human brains. Neural networks rely on training data to learn and improve their accuracy, once the algorithm is built it can be used in classification or clustering tasks. The research [45] shows Deep Neural networks (DNNs), Convolutional Neural networks (CNNs), and Recurrent Neural networks (RNNs) are the three most popular deep learning models used in the area of sentiment analysis. The research uses those methods on eight data sets to check the performance of the models. word embedding and TF-IDF is used in the feature extraction process. In general, word embedding is a more robust technique than TF-IDF not only in the accuracy but also takes less computational time. The CNN model was found to offer the best tradeoff between the processing time and the accuracy of results. Although the RNN model had the highest degree of accuracy when used with word embedding, its processing time was 10 times longer than that of the CNN model. DNN is a simple deep learning model that has average processing times and yields average results.

Advantage

- They can capture the information contained in large amounts of data and build incredibly complex models
- Good performance in noisy data

Disadvantage

- Long time to train
- High memory usage

2.3.3 Hybrid Approach

The hybrid approach of sentiment analysis integrates both lexicon and machine learning methods for polarity detection. It derives accurate results from machine learning (statistical approaches) and stability from a lexicon-based approach [46]. Different research focuses on performing a hybrid approach to perform sentiment analysis. The researchers in [47] adopt a hybrid approach for enhanced Twitter sentiment. The research will take into account the advantage of both approaches (lexical and statistical). The research uses the most frequent and general-purpose SentiWordNet (SWN) lexicon resource in the first phase to extract feature vectors and SVM is trained on SWN-based generated feature vectors. When extracting features negation handling is done through a shifting approach rather than the traditional reverse polarity approach to increase performance. The finding shows substantial improvement in using this approach than the traditional one using macro-averaged recall and F1 score.

- The hybrid approach as examined in the latest research efforts can give us good classification performance.

2.4 Feature Extraction Methods

Feature extraction is an important process in performing sentiment analysis. The procedure entails converting raw data into numerical inputs required by the machine learning algorithm. To extract features from textual data, Bag of Words, TF-IDF, and Word2Vec can be used [48].

2.4.1 Bag of words

A bag of words is the simplest form of text representation in numbers. Using a Bag of words sentence can be represented as a bag of words vector.

The disadvantage of a Bag of words is it will create a vocabulary from all unique words. So, if the new sentence contains new words the vocabulary size will increase. So, they are not applicable for larger datasets as they degrade performance and create a sparse matrix vector containing many zeros.

2.4.2 TF-IDF

Term frequency-inverse document frequency is a numerical metric, that measures how important a word is to a document in a corpus of documents. In the case of sentiment analysis, it can be used to extract the most frequently used word in all of the comments or reviews. Unlike the Bag of words, in which the word with higher frequency becomes dominant in the data in TF-IDF, the frequency of the words is rescaled by considering how frequently the words occur in all the documents.

Mathematically, TF-IDF can be represented using:

$(\text{Tf\_idf})_{t,d} = \text{tf}_{t,d} * \text{idf}_t$, The term frequency-inverse document frequency of a term t in document d is the product of the term frequency of that term in the document with the inverse document frequency of the term.

The primary issue with both TF-IDF and bag of words is that they do not consider the grammatical structure or context of the term. Word embedding techniques are required to identify the commonalities between different terms in our corpus.

2.4.3 Word Embedding

In word embedding, words with comparable meanings will have similar representations. For Example, the words “አስደንጋጭ” and “አስፈሪ” will have similar representations or their embedding can be represented very close to each other in a 2D plane. The word embedding feature can help the classification model to model the semantics of a word [49].

The main idea of word embedding is to represent each word using a densely distributed representation. They include a set of feature selection methods or language models that map a textual word into its dense and low-dimensional vector representations. This can make the model performs well than sparse vectors. Word embedding is used in different research and shows good results. The research in [45] uses word embedding and term frequency-inverse document frequency (TF-IDF) to convert raw data into numerical vectors after preprocessing. The result shows word embedding is a more appropriate technique than TF-IDF for performing sentiment analysis in most datasets. Different word embedding techniques are used to learn word embedding from text data.

2.4.3.1 Word2Vec

Word2vec is a technique introduced by Thomas Mikolv and his team at Google in 2013 [50]. The algorithm uses a neural network model to learn word associations or embedding from a large corpus of text then the algorithm can predict between every word and its context. So, words occurring in similar contexts are related. Two algorithms are introduced as part of word2vec known as Continuous Bag-of-words (CBOW) and Continuous Skip-Gram Models. The CBOW model tries to predict the current word (target) based on the context of surrounding words. In the case of the Skip-gram model, the reverse logic works. So, we will take a series of words and the model will try to predict the context. The following table shows the training samples used by Skip-gram to predict the context of a given word. In the example window size of two is used. This means we will pick two words behind a target word and after it. The underlined word represents our target word.

Table 2.2: How word2vec works

Input Sentence	Training samples
ከጤና ሚኒስቴር የተላለፈውን መመሪያ ሁላችንም የማከበር ግዴታ አለብን::	(ከጤና,ሚኒስቴር),(ከጤና,የተላለፈውን)
ከጤና ሚኒስቴር የተላለፈውን መመሪያ ሁላችንም የማከበር ግዴታ አለብን	(ሚኒስቴር ,ከጤና), (ሚኒስቴር , የተላለፈውን), (ሚኒስቴር , መመሪያ)
ከጤና ሚኒስቴር የተላለፈውን መመሪያ ሁላችንም የማከበር ግዴታ አለብን	(የተላለፈውን,ከጤና),(የተላለፈውን,ሚኒስቴር),(የተላለፈውን,መመሪያ),(የተላለፈውን,ሁላችንም)
ከጤና ሚኒስቴር የተላለፈውን መመሪያ ሁላችንም የማከበር ግዴታ አለብን	(መመሪያ,ሚኒስቴር),(መመሪያ,የተላለፈውን),(መመሪያ,ሁላችንም),(መመሪያ,የማከበር)

The training word pairs will be used for training the skip-gram model. The input will be the main word in one hot encoding and we will feed it to a shallow neural network this network will learn the nearby words and try to predict the most probabilistic word as an output.

Skip gram works well with a small amount of the training data and it can represent well even rare(less) frequent words or phrases, whereas CBOW is several times faster to train than the skip-gram, with slightly better accuracy for the frequent words [50].

So, we can conclude feature extraction is an important process in sentiment analysis. The classifier needs the output of the feature extraction process, which is a numerical vector this can be very helpful in dimensionality reduction and to get the syntactic as well as semantic structure of our dataset.

2.5 Evaluation Mechanism

Sentiment analysis is mostly used to classify text into a class (positive or negative). Such type of classification problem uses evaluation metrics to understand to what extent our methodology performs well. The most commonly used type of evaluation metrics is Accuracy, Precision, Recall, and F1-Score.

Accuracy

Accuracy is the percentage of correct predictions made by a classifier.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision

Precision talks about how precise/accurate your model is out of those predicted positive, and how many of them are actually positive. It is good for measuring whether or not our costs of false positives are high or not. In sentiment analysis "False positive" refers to a condition to classify a text as negative while the test is positive.

$$\text{Precision}= \frac{TP}{TP+FP}$$

Recall

Recall calculates how many of the actual positives are captured by our model as positive. It is good for measuring whether or not our costs of false negatives are high or not. In the case of sentiment analysis "False negative" refers to a condition to classify a text which has a positive polarity as negative.

$$\text{Recall}= \frac{TP}{TP+FN}$$

F1-Score

F1-Score is needed to balance between precision and recall.

$$\text{F1}= 2 * \frac{Precision*Recall}{Precision+Recall}$$

2.6 Related works sentiment analysis on covid19

The covid-19 impacts different sectors worldwide as we have discussed in the introductory section. Different countries try to mitigate the influence of this pandemic by making efficient strategies that must be undertaken by gaining input from citizens. Machine learning and sentiment analysis will play a major role in such cases, as many people try to share their ideas on different social media so understanding the discourse is very important. The prior trends show such techniques are used to control different viral infections and mitigate the consequence. Several study articles have shown that many outbreaks and pandemics may have been quickly contained if specialists had considered social media data by providing prior studies on flu and influenza control [51].

Next, we attempt to review work that has been done about the covid-19 epidemic.

2.6.1 Covid-19 sentiment analysis for English language

The research done in [18] focuses on gaining the public attitudes towards covid-19 vaccination in UK and US by using sentiment analysis. The problem they try to address is the concerns of vaccine skeptics and develop the required public trust in immunization to realize the goal of generating herd immunity. This needs continually monitor and a better understanding of public sentiments to develop baseline levels of confidence in vaccines. To achieve this the objective data is collected from both Twitter and Facebook using Twitter API and the Crowd Tangle platform respectively. They consider the time frame from March 1 to November 22, 2020, the filtered dataset was initially preprocessed (For example: remove links, hashtags and stop words) and then for thematic sentiment analysis, a new hierarchical hybrid ensemble-based AI model was built. using a rule-based ensemble method, two lexicon (rule) based approaches, VADER and Text Blob, are merged with a pretrained DL-based model, Bidirectional Encoder Representations from Transformers (BERT). They find over a 9-month study period, the average positive, negative, and neutral sentiments were at 58%, 22%, and 17% in the United Kingdom, compared to 56%, 24%, and 18% in the United States, respectively. Concerns about vaccination safety, economic viability, and corporate control were highlighted with public positivity about vaccine development, effectiveness, and trials. They state as limitation users are known to differ in their usage and preferences for social media platforms based on their sociodemographic (e.g., age, socioeconomic status, and political affiliation), and they indicate that their statistics may not be representative of the broader population of both countries.

The research in [52] was done by looking into the few studies undertaken about general public responses in Canada and also prior work to determine the sentiment of an overall text rather than capturing opinions toward COVID-19 specific aspects chosen by domain experts and exploiting lexicon built-in general domains. They use a public Twitter dataset about the covid 19 pandemic collected by Chen et al using numerous COVID-19–related keywords such as "coronavirus," "COVID-19," and "pandemic." They consider the time frame from January 21 2020 to the end of May 2020. They first discovered topics in COVID-19–related tweets using a widely used topic modeling approach, latent Dirichlet allocation (LDA) because of its simplicity and popularity. The topics generated by LDA were interpreted and labeled by two public health experts. To capture sentiment revealed in tweets toward important aspects of COVID-19, they used ABSA (aspect-based sentiment analysis). They include public health interventions or issues associated with COVID-19, such as "social-distancing," "reopening," and "masks." and they investigate people's opinions (positive and negative) toward these aspects. ABSApp tool is used which is a weakly-supervised ABSA system for sentiment extraction. COVID-19 transmission was the most talked-about topic in both Canada and the United States. In both nations, there was discussion about the initial outbreak in Wuhan and US President Donald Trump's response. They collected 545 aspects and 60 opinion phrases after training the data with ABSA with humans in the loop. COVID-19 transmission was the most talked-about topic in both Canada and the United States. In both nations, there was discussion about the initial outbreak in Wuhan and US President Donald Trump's response. They collected 545 aspects and 60 opinion phrases after training the data with ABSA with humans in the loop. The findings revealed negative feelings about the general outbreak, misinformation, and Asians, as well as positive feelings about physical distance. As a limitation, they state a small set of Twitter datasets was used due to the location information being limited compared to the whole dataset, and the geotagged tweets data set comprises statements from a nonuniform subsample of the population.

The research in [53] focuses on identifying the Indian public's perspective on anxiety, stress, and trauma during Covid-19. The aim of the study is that governments and the officials who are policymakers need to understand the issues that cause anxiety, trauma, and stress among the general public during the pandemic time so that appropriate recovery measures can be taken. Tweets from seven months are used in the research. Sentiment analysis and Topic modeling are used in the study; the sentiment analysis is undertaken because the objective of the study was to analyze the tweets that talk about COVID-19 and the anxiety, trauma, and stress caused by it, It is not possible to assume that all the tweets must be in a negative tone. Topic modeling is used to figure out the top issues, people discuss while posting about the stress, trauma, and anxiety caused by COVID-19. Even when discussing the stress, worry, and trauma triggered by COVID-19, they discovered that the majority of tweets were neutral. The two major parts of the COVID-19 that created worry, anxiety, and trauma among Indian civilians were death and lockdown. The limitation stated the research focus on the generalized perception of the Indian populations but they recommend studying how the perceptions differ between prominent cultures and subcultures.

Dynamical analysis is important for authorities to understand when and where people would experience mental health issues. The research undertaken in [54] focuses on detecting topic and sentiment dynamics as a result of the COVID-19 pandemic using social media. The task is broken down into three steps in the proposed framework. They partition the tweets into different subjects using a dynamic topic model, then determine the sentiment polarity of each tweet, and then summarize the sentiment polarity distribution of each topic using a sample of 13 million tweets about COVID-19 collected over two weeks. During the study period, they discovered that positive sentiment has a larger ratio than negative emotion. When they look at the topic-level analysis, they discover that distinct components of COVID-19 have been discussed frequently and have similar sentiment polarities. Positive sentiment dominates some issues, such as "remain safe at home." Others, such as "people's death," frequently reflect a negative attitude. Overall, the proposed framework shows insightful findings based on the analysis of the topic-level sentiment dynamics.

2.6.2 Covid-19 sentiment analysis for non-English language

The views about the covid-19 pandemic can vary from country to country. The Weibo social platform is one of the famous platforms used in the Chinese community.

The research done in [55] focused on analyzing the negative sentiment of the public using the Weibo platform. The research they perform try to understand the public negative sentiment on social media by performing thematic analysis to identify the reasons for people’s negative post and investigate the trends of sentiment development and the underlying major themes to understand public concerns. 999,978 microblogging posts from Jan 1, 2020, to Feb 18, 2020, are analyzed. BERT unsupervised machine learning algorithm is used to classify the sentiment of Weibo posts into positive, neutral, and negative categories and to analyze the trend of posts. TF-

IDF algorithm to extract topics of posts with a different sentiment. The research finds that 56.2% of posts are neutral, while there is more positive sentiment (27.4%) than negative sentiment (16.4%). Before 19 January, the number of posts about COVID-19 is quite stable. As of 20 January, there is an increase of 25.81% and 33.03% in negative posts and total posts, respectively, and a decrease of 16.30% in positive posts as compared to that of 19 January.11 key topics related to negative sentiment along with the corresponding themes and frequency of occurrences in Weibo posts are also discussed.

Dutch General Public Reaction on Governmental COVID-19 Measures and Announcements using Twitter data is performed in [56] research. Covid-19 related tweets are extracted using keywords. The tweet's polarity is automatically assigned using the pattern (Dutch sentiment lexicon) python library. The research finds the sentiment on the COVID-19-related topic tends to be more negative than that of Dutch tweets on average. They also found that national press conferences about governmental measures and announcements (e.g., the first national press conference about the lockdown policy) can influence public sentiment. Public attitudes towards Specific policies are measured by taking two cases; social distancing and wearing face masks. For both cases, the annotation scheme included three labels from extracted tweets: Supports, Rejects, and Other. The analysis showed that people widely supported social distancing when the measure was announced in March, then support declined until June and increased again recently analysis regarding face masks, the analysis showed that the public widely rejected the initial government position that face masks are useless for the general public.

One of the impacts covid-19 made is the market area. The research done in India [57] focused on sentiment analysis and prediction of the Indian stock market. The aim is to identify the most suitable technique for sentiment prediction based on accuracy with a special focus on the period which covers the outbreak and spread of the Covid-19 pandemic. The data is collected from RSS Feed, Forum discussion, Twitter, and News portal. After data is preprocessed Bag of words, TF-IDF, and N-grams are used separately for selected 6 ML algorithms which are Decision Tree, KNN, Logistic Regression, Naïve Bayes, Random Forest, and SVM. While comparing the Bag-of words technique with TF-IDF technique for the given data, it was observed that the Bag-of-words technique has shown higher classification accuracy in terms of sentiment prediction. Both Logistic Regression and Support vector machine algorithms have indicated the highest classification accuracy of 78% for sentiment prediction.

The research [58] tries to collect and prepare a dataset that classifies an individual's awareness of following precautionary measures that are provided by the government during the quarantine period in Saudi Arabia. The methodology they follow is data collection with Twitter API, the data is collected from each Saudi Arabia region (south, north, east, west, and middle), then proper preprocessing is undertaken on the data based on language structure. Then the tweets are annotated into three categories; positive, negative, and neutral. Then feature extraction is performed using N-Gram and TF-IDF then the machine learning model is used to classify the sentiments. They use SVM, Naïve Bayes, and K-Nearest Neighbor classifiers. The result shows that bigram TF-IDF with the SVM classifier produced the highest accuracy of 85%, which outperformed KNN and Naïve Bayes. the south region showed the highest level of awareness at 65%, while the middle was the lowest among the regions. From this research, the feature extraction is N-gram with TF-IDF such type of feature selection doesn't get the semantic structure of the sentence as well as it has larger feature dimensions which decrease the performance of the classifier. So, it will be better to apply feature selection and parameter tuning while building machine learning models to increase performance.

2.6.3 Covid-19 sentiment analysis for Amharic Language

In our local language, there are many works done focusing on sentiment analysis but in this covid-19 domain, there are still additional works to be undertaken to know more about the public attitudes towards this pandemic.

Hiwot Wonago [59] created a model that filters information on social media using Sentiment analysis. The "information" that was filtered refers to any Toxic online content used to tackle politically and socially sensitive content, and prevent illegal or unsuitable social media content from being accessed. The data set was prepared and labeled into four classes politically offensive, Socially-offensive, Religiously-offensive, and non-offensive labels. The results showed that the support vector machine performs comparatively better with the word2vec approach.

The research done in [60] also tries to perform sentiment classification using semantic feature extraction based on the Word2vec approach for Amharic language text in a political domain. Word2Vec and TF-IDF were used to learn the word representations as a candidate feature vector then two machine learning algorithms are used to perform sentiment classification. Gradient-Boosting Tree (GBT) and Random Forest machine learning algorithms are used to train and test in the Apache Spark platform. The experiment shows feature extraction using the Word2Vec technique performs better in the GBT classifier achieving an accuracy of 82.29%.

The research [61] tries to explore Amharic sentiment analysis using social media texts. To achieve this task, they propose “ASAP” a social network-friendly annotation tool using Telegram bot that is for annotating the labels used for sentiment classification. The data is collected from Twitter using Twitter API. Around 9,400 tweets are collected and labeled by 3 telegram users. The sentiment analysis is then performed using machine learning. Then they discover The FLAIR deep learning text classifier outperforms conventional supervised classifiers by using network embeddings calculated from a distributional thesaurus.

2.7 Research gap and Summary

In general, From the literature review the overview of sentiment analysis, the levels and approaches that are used to perform sentiment analysis, and the effect of feature extraction in making sentiment analysis are reviewed. State-of-the-art techniques used in sentiment analysis and their efficiency are also discussed.

From the analysis of related work, we understand most of the research on sentiment analysis of social media to assess people’s attitude toward covid-19 has focused on the English language and European languages. And in some reviewed research works the researchers suggest users' usage and preferences for social media platforms are known to vary depending on their sociodemographic (for example, age, socioeconomic status, and political affiliation). So, it is important to see and examine the people’s opinion towards the directions and reports made through the Ministry of Health on local languages like morphologically reach languages such as the Amharic language. To the best of our knowledge, there is no work done related to this domain (sentiment analysis of people’s attitudes on covid19 using social media). The paper tries to review some efforts that were taken in the area of sentiment analysis of Amharic texts. The machine learning model will highly depend on the domain that is being applied in Natural language processing. Learned models often have poor adaptability between domains or different text genres because they often rely on domain-specific features from their training data [35]. Because we can’t apply the model applied in another domain easily to our domain, we have to create a dataset related to the covid-19 area. The main aim of this research is to fill the gap observed by collecting data sets specific to covid-19 in the Amharic language and assess the people's attitudes towards this pandemic that can be used mainly by public health authorities in making effective strategies.

CHAPTER THREE

DESIGN AND METHODOLOGY

3.1 Introduction

In this chapter, the overall design that is used to perform sentiment analysis on Amharic text towards to covid-19 reports and directions given by the Ministry of Health were discussed. The architecture that shows the high-level flow of the system was graphically presented. The research use state of the art machine learning algorithm to perform sentiment analysis. The reason behind this is the availability of data is considered the key to construct a machine learning model or data-driven real-world systems [62]. Since the data that is used to perform the sentiment analysis is not much enough thus using well-known and tested models according to their capabilities is considered as a solution. In addition, employing a hybrid approach is difficult due to the limitation of available rule-based sentiment lexicons on local languages like SentiWordNet or Vader makes it difficult to test the hybrid approach by combining it with machine-learning algorithms to come up with a new novel method. The process that we follow was comments which shows expression on covid-19 are applied as an input for the system at sentence level then the system classifies whether each sentence is a positive or negative sentiment as an output based on the training performed. In generating such types of outputs, several steps are followed as it is elaborated in the architecture. Then each component that interacts in the system was discussed. The output generated by the system can be very helpful for government officials to track the tones that are discussed on social media which can be very helpful in decision making.

3.2 Architecture of the classifier model

As mentioned at the beginning of the chapter a general structure of machine learning was adopted to extract and perform sentiment analysis. The sentiment classification system is generally composed of data collection, the data filtering module, Morphological analyzer, Feature extractor, Machine learning model builder, and finally the classification task. The diagram that shows the architecture was presented below then each component's functionality discussed in section 3.3.

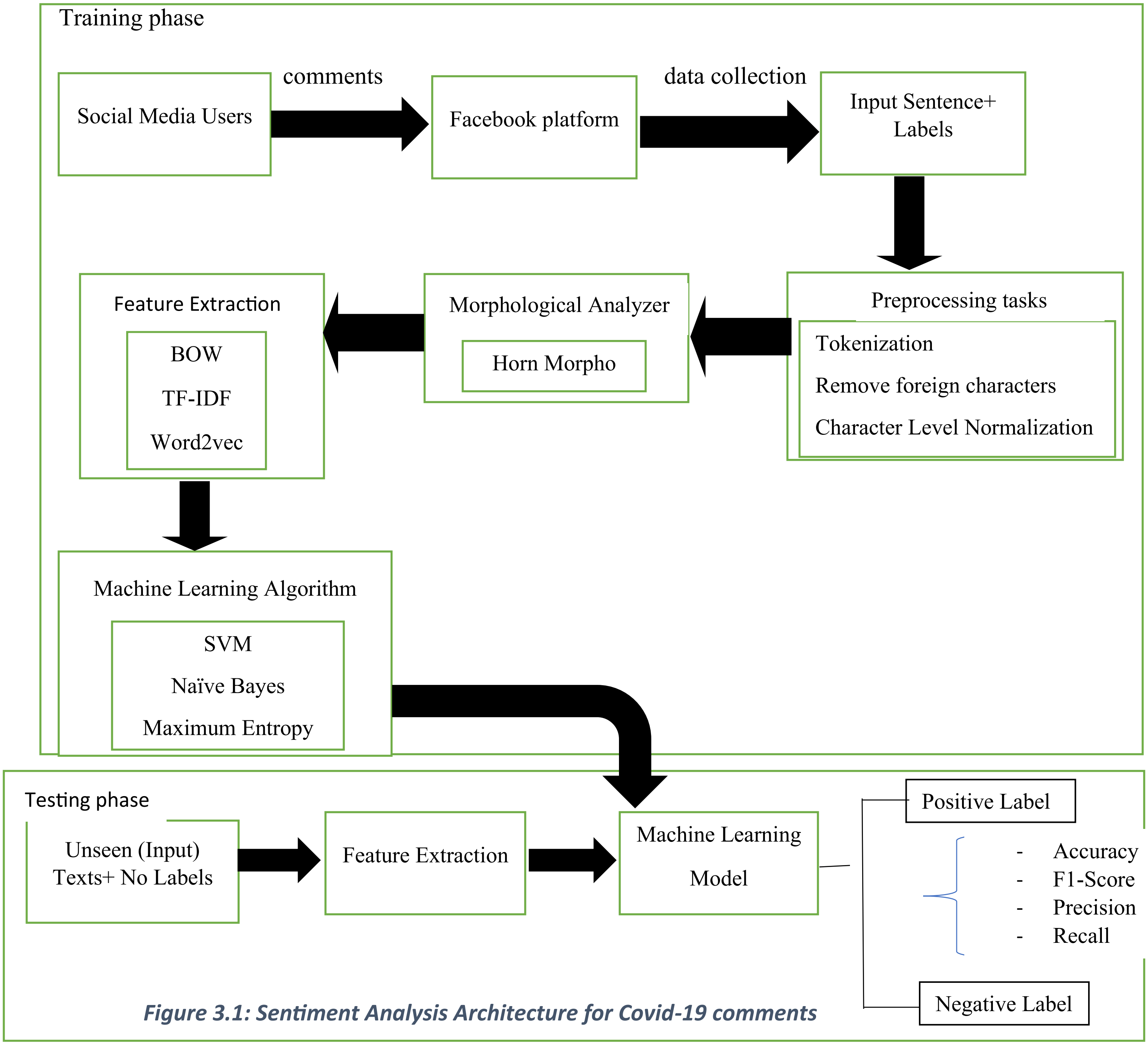
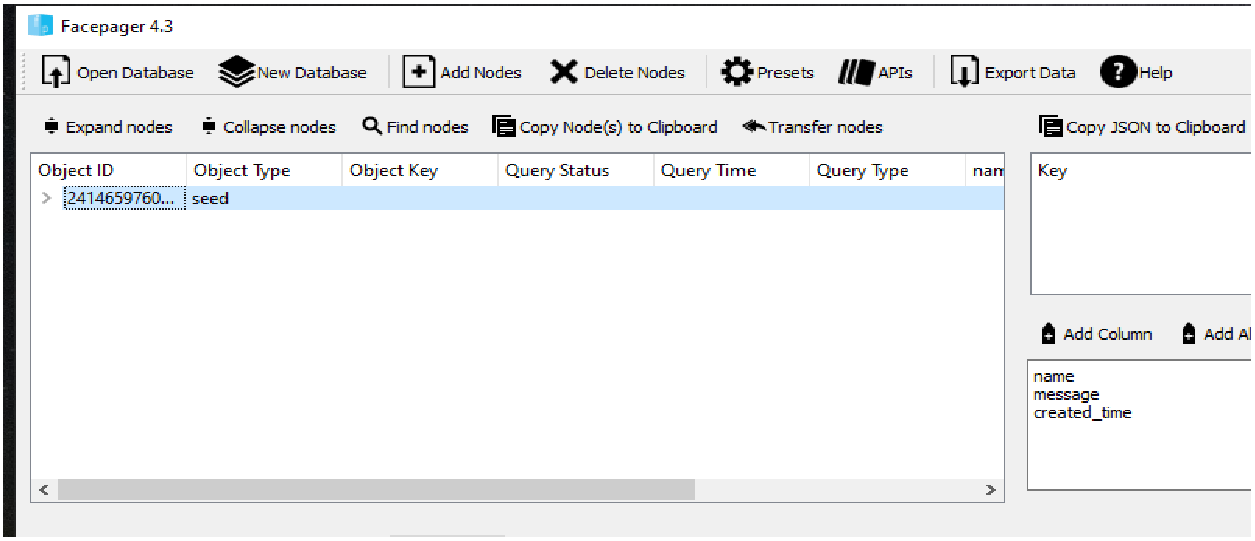


Figure 3.1: Sentiment Analysis Architecture for Covid-19 comments

3.3 Components of Sentiment Classification System

3.3.1 Data Collection

The data collection was mainly done on the Facebook platform because of the higher community on Facebook compared to other social media networks in our country. Since, we are dealing with people's comments about covid-19 the data is without proper formatting and alignment, this makes our data unstructured. The main source for getting the data was the Ministry of Health (MOH) official Facebook page. Facepager [63] tool was used to extract the data from the platform, the tool is designed for fetching data from Facebook, Twitter, YouTube, and other JSON-based APIs.



Object ID	Object Type	Object Key	Query Status	Query Time	Query Type	nan
2414659760...	seed					

Key

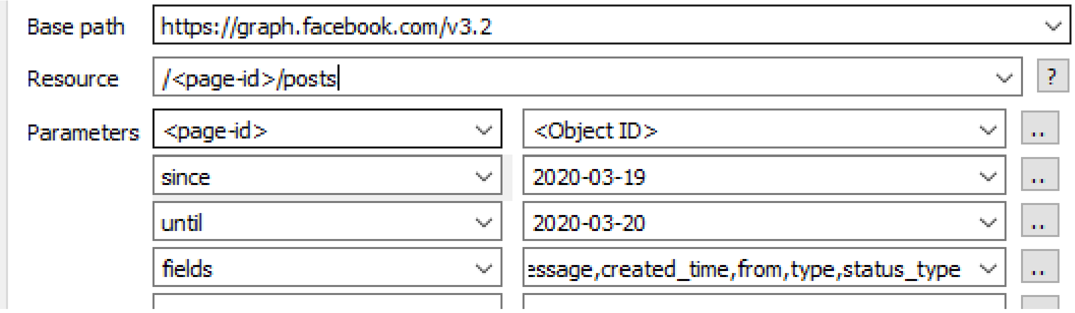
name
message
created_time

Figure 3.2: Face Pager View

Steps followed by Facepager to access comments.

1- Click Add Node button to add one node in the face pager for analysis purposes

2- Get the Facebook id of the Ministry of health from https://findmyfbid.in/ website. i.e., In this case, the no 2414665…. refers to the id of the page



Base path https://graph.facebook.com/v3.2

Resource /<page-id>/posts

Parameters	<page-id>	<Object ID>
	since	2020-03-19
	until	2020-03-20
	fields	ɜssage,created_time,from,type,status_type

3- the resource page-id/posts are used to access the posts from the page within the required time frame

4- the post-id/comments are used to access the comments from each post. Each post will have its id.

5- Then by using the export data feature the data is converted to CSV format.

3.3.2 Preprocessing tasks

The most important thing in dealing with most NLP application including sentiment analysis is data preprocessing, because we are dealing with unstructured textual data which contain noisy, wrong, or inconsistent data such data needs extensive data cleaning. by carrying out this task, we transformed the data into a basic form that makes easy to work. This task can increase or improve the classifier efficiency and accuracy. Next, we try to discuss the main techniques used to perform data preprocessing.

3.3.2.1 Tokenization

Tokenization is the process of splitting a piece of text into smaller units called tokens. The tokenization process can be done at word, character, or sub-word levels. This process can be very useful to understand the meaning of the text and prepare a vocabulary. The most common way of getting tokens is to use space between words as a delimiter but this depends on the language behavior. For example, in the Chinese writing system, there is no whitespace between phrases [64]. So, splitting based on white space may not work well like in other languages. In the Amharic language to separate a sentence, we usually use space as a delimiter to find tokens or words, and sentences are separated by a full stop (: :).

The below table shows how the tokenization algorithm works.

Table 3.1: Tokenization Algorithm

Input	Sentence
Output	Tokens (words)
Process	1) Read each sentence
	2) Break the sentence into words using a delimiter (space)
	3) Append each word into a list
	4) Return a list of words

3.3.2.2 Data Cleaning

Data cleaning is the process of preparing data for analysis by removing unnecessary data that can hinder the process or provide inaccurate results. The first thing we deal with in data cleaning is removing unwanted characters from our dataset. In Amharic language we get different characters like question marks (? ጥያቄ ምልክት), (፤ ድርብ ሰረዝ), (፣ ነጠላ ሰረዝ), (! ቃል አጋኖ). These characters don’t have any relevance to sentiment analysis tasks. The most frequently used stop-words in Amharic documents are “እና”,” ወደ”,” እኔ”,” ውስጥ”,” እነሱ”, ‘‘ነው”, and ‘‘ነበር”. This and additional stop words should be identified and removed from the dataset. The python Regex (Regular expression) is used to perform the preprocessing task.

The below table shows how the data cleaning algorithm works in the system.

Table 3.2: Data Cleaning Algorithm

Input	Words (sequence of characters)
Output	Filtered words
Process	1) Read each word from the list
	2) If each word is punctuation, stop word, digits, or irrelevant Unicode character perform replacement.
	3) Return clean words

3.3.2.3 Normalization

Normalization is helpful in reducing the number of unique tokens present in the text, removing the variations in a text, and also cleaning the text by removing redundant information. In Amharic language, we have different characters with the same sound but having different orthographic structures as mentioned in the Literature review section. Such representation can create redundancy in our system so they must come into one form.

The below table shows how the Normalization algorithm works.

Table 3.3: Normalization Algorithm

Input	Words
Output	Normalized Characters
Process	1) Read each word
	2) Break the words into characters

	3) If the character is ኀ, ሐ then replaces with ሀ Else if the character is ሠ then replaces with ሰ Else if the character is ዓ, ዐ then replaces with አ Else if the character is ፀ then replaces with ጸ End if
	4) Return the normalized word

The Normalization process can also include converting a word into its base form. This will be discussed in the next section in morphological processing.

3.3.3 Morphological processing

Morphological processing tries to understand the internal structure of words by identifying the root or stem form of the word. For example, for the word "ተማሪዎቹ" the root word is "ተማሪ". Morphological processing is essential for increasing the performance of the classifier because the classifier only deals with root words extracted by a morphological analyzer which can help us to minimize the representation of words.

Horn Morpho [65] was used to perform morphological processing. It is an open-source morphological analyzer for Amharic, Tigrigna, and Afan Oromo languages. In Mikael gasse‘s morphological analyzer the stem is produced by eliminating inflectional terms from words (prefix and suffix). The stem is processed by the stem Finite State Transducer (FST) which extracts the root, with various grammatical properties.

The below table shows how morphological processing works.

Table 3.4: Morphological Processing Algorithm

Input	Words
Output	root word
Process	1) Read each word
	2) Pass the word into Horn morpho

	3) Select the root word from the segmented word generated by Horn morpho 4) Return the word

3.3.4 Feature Extraction

The feature extraction process is marked as the most important step in the area of machine learning. The process involves changing the textual representation into numerical representation for it to be used by Machine learning or any other solution. So before applying machine learning we have to undertake feature extraction. There are different techniques used to describe the properties possessed by the data by mapping higher dimensional data onto a set of low dimensional potential features. In this thesis work, TF-IDF and Word2vec were used to extract features.

3.3.4.1 BOW

The method tries to represent each word as a list of word vectors based on taking how many times (occurrence) appears within the corpus. It doesn't take into account any information about the structure or order of the sentence in the feature representation phase. This numerical representation allows the system to understand the data and process easily. In the bag of words, the word usually represented using 0 and 1 to track it is presented in the sentence or not.

3.3.4.2 TF-IDF

This method tries to normalize the frequency of a given word based on other documents. Unlike a simple bag of words that represents only the existence of words in a document (corpus), TF-IDF tries to give how important a given word is in a corpus by giving a score for each word. The score of a word can be calculated using

$$\text{TFIDF (word, doc)} = \text{TF (word, doc)} * \text{IDF (word)}$$

For example, for the word "ኮረና" there are two calculations that should be performed this will and as an outcome, we get a matrix representation for the word. First, the word frequency in the whole corpus is to be calculated and the inverse document frequency of a word is as follows.

$$\text{TF (word, doc)} = \frac{Frequrncy\ of\ a\ word\ in\ the\ doc}{No\ of\ the\ words\ in\ the\ doc}$$

$$\text{IDF (word)} = 1 + \frac{No\ of\ docs}{No\ of\ docs\ with\ word} \quad \text{or log} \left(\frac{No\ of\ docs}{No\ of\ docs\ with\ word}\right)$$

3.3.4.3 Word2-vec

Techniques like TF-IDF and bag of words don't consider the semantic meaning of the words or the contextual meaning of the words present in the vocabulary. To investigate the semantic relation embedding, which uses a similar type of embedding to represent words with related meanings. The context of a word can be detected by incorporating information about preceding and following words in the vector that represents a given occurrence of a word. This allows us to achieve significantly better outcomes in natural language processing tasks. In this work Word2vec using skip-gram is adopted to extract features. The output is a fixed vector representation for each word.

Table 3.5: Word2Vec Algorithm

Input	Word
Output	Set of vectors
Process	1) The comments will be divided into word level and morphological processing will be performed at word level
	2) The word will be given as one-hot vectors, the vector length will be similar to the vocabulary length. The vector is filled with 0's except the index that represents the word we want to represent, which is assigned 1.
	3) The one-hot encoded vector will go through a hidden layer whose weights are the word embeddings. The weights are adjusted by minimizing the loss function and two edges exist between nodes(words) only if their corresponding vocabulary co-occurs in a window of length K, where K represents the size of the window.
	4) Outputs the candidate keywords from the vocabulary

- In this research both the word2vec and TF-IDF including BOW are used to see and compare their efficiency to check how feature extraction affects the classification works.

3.3.5 Machine Learning algorithms

There are a lot of state-of-the-art classification algorithms available for making sentiment analysis including machine learning as well as deep learning models. Choosing the right type of algorithm depends on the dataset used to perform the analysis. In this paper due to the limitation of data size SVM, Maximum Entropy, and Multinomial Naïve Bayes are used to train sample data and perform classification tasks.

Due to the latest survey [66] that focuses on sentiment analysis the most commonly used Machine learning algorithms are SVM and Naïve Bayes with 41% and 35 % respectively.

Next, we try to see how machine learning works by discussing the high-level algorithm representation of the models.

3.3.5.1 Support Vector Machine

SVM algorithm is used mainly in textual classification works and it proves one of the most powerful learning algorithms to perform classification. The first benefit that we get from SVM is it can handle high dimensional input features because the algorithm does not depend on several features due to overfitting protection by choosing a specific hyperplane that can separate the data in feature space [67].

SVM algorithm tries to get (X, y) data as an input where X refers array of input with m features and y refers array of class labels. Then the algorithm starts the training process from the data ingested and tries to maximize the output by updating weight and calculating the likelihood of each category. Basically, SVM tries to construct a hyperplane in multidimensional space to separate different classes. To separate lines better, SVM uses support vectors which are data points closer to the hyperplane. The distance between the support vectors and the line is called the margin. If the margin is maximized, we can better classify classes.

So, the major aim of SVM is to select a hyperplane with the maximum possible margin between support vectors in our dataset.

```
Input: X,y where x refers array of input with m features
                          Y refers array of class labels
Output: Optimized Machine learning model
Process:  train_svm (X,y, no_of_runs)
initialize: learning_rate= Math.random();
for learning_rate in no_of_runs
     error=0;
     for i in X
   if(y[i]*(x[i]*w))<1 then
update: w = w + learning_rate * ((x[i]*y[i])*(-2*(1/no_of_runs)*w)
else
update: w= w + learning_rate * (-2*(1/no_of_runs)*w)
end if
end for
```

Algorithm 3.1: pseudo code of SVM training

3.3.5.2 Naïve Bayes

Naïve Bayes classifier is used in the study based on the inference from the literature review undertaken due to its simplicity and no assumption drawn between the features used as well as it is fast algorithm [39].

For each class, Nave Bayes predicts membership probabilities, such as the likelihood that a given record or data point belongs to that class. The most likely class is the one with the greatest probability.

```
Input: Sentence level comments
Output: Optimized Machine learning model
Process:  TrainMultiNomialNB (C,S)
 V ← ExtractVocablary (S)
 N ← CountSentence (S)
for each  $c \in C$
do $N_c$ ← CountSentenceInClass (S, c)
    prior[c] ← Nc/N
    $text_c$ ← Extract all Text of SentenceInClass (S,c)
    for each t ∈ V
    do conditionalprob[t][c] ← $\frac{Tct+1}{\sum_{t'}(\mathrm{Tct}'+1)}$
return V, prior, conditionalprob
```

Algorithm 3.2: pseudo code of Naïve Bayes training

3.3.5.3 Maximum Entropy

It is also known as a conditional exponential classifier or Maxent classifier. Like the Naïve Bayes algorithm, it doesn’t make assumptions about the relationships among features. It tries to estimate the conditional distribution of the class label C given a document d to maximize the entropy of the system by using the following exponential form [68].

Max Entropy is used even if it follows the same working principle as naïve bayes it can give us some advantages like non independent assumption, which is found often in the text classification problem but it needs a very long time to train data compared to Naïve Bayes [69].

$$P_{ME}(c|d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda i, c\ fi, c\ (d, c)\right)$$

where Z(d) is a normalization function, $f_{i,c}$ is a feature function for the feature f_i and the class c, and $\lambda_{i,c}$ is a parameter for the feature weight to make sure that the observed features match the expected features in the given set.

$$f_{i,c}(d, c') = \begin{array}{l} 1, n_i(d) > 0 \text{ and } c' = c \\ 0, \text{otherwise} \end{array}$$

3.3.6 Classification Task

The comments extracted were used as a dataset in the analysis. In the training step, we used our data to incrementally improve our model‘s ability (tune) to predict whether the posted comments contain positive, negative, or neutral towards the direction given by the ministry of health. For training, we have to use a machine learning model for the feature extracted comment data with their labels (Positive, Negative). Then the machine learning tries to understand the patterns that are required to classify the comment to our target (output). Then we will feed unseen data to our model to predict the required label. The unseen comments will be represented as features and then they will be checked by the machine learning model. The Testing phase has the following Component.

i) Testing Set: The testing phase consists of data that are used to check to what extent the machine can classify comments to their respective category. In most cases, 80% to 20% or 75% to 25% is used for training and testing respectively.

ii) Preprocessing task: testing data will be filtered by removing unnecessary characters, numbers, or special characters to make things easy for the classifier.

iii) Feature Extractor: in this phase, each word used in the testing phase will have a vector representation similar to the steps followed in the training process.

iv) Sentiment Classifier: In this phase, the pre-trained supervised machine learning models are used to predict the given test set to their corresponding sentiment.

v) Testing (Evaluation): In this phase, we measure to what extent the trained machine learning model can categorize or predict the comments as Positive or Negative correctly using quality metrics.

3.4 Summary

The main output of the methodology section is to create artifacts since we are using the design science approach. The artifact will be used to design the best sentiment analysis model that predicts people‘s expression on covid-19 by using well-known machine learning models and tools. The process that is followed includes data collection from the Facebook platform followed by extensive data preprocessing tasks like data filtering, tokenization, and text Normalization to minimize the data representation and increase the efficiency of the system. Once the data preprocessing is done then the system performs feature extraction using TF-IDF and Word Embedding using word2vec and builds a machine learning model using SVM, Naïve Bayes, and Maximum Entropy. Finally, the model's effectiveness should be tested by using sample data set to check the machine's effectiveness and rigor by using available metrics like accuracy, precision, recall, and F1 score. The finding can be which one of the combinations (machine learning model and feature extraction) method is better for our objective.

CHAPTER FOUR

EXPERIMENTATION AND EVALUATION

4.1 Introduction

In this chapter, the experimental result of the system is discussed including the procedures used for data collection, pre-processing, and the classification result based on different techniques and their efficiency to what extent the method achieves the objective was presented. The main challenge in performing this experiment is the lack of efficient and quality data that focuses on the covid-19 related posts. The main reason for this is the way of writing comments on the channels varies from person to person. Since the research focus on Amharic text, finding Amharic only text is challenging. Another challenge is the data source that is used for experimental purposes, based on some security constraints on the official health institute's social media page we are highly depend on the Ministry of Health (MOH) official page.

4.2 Data Preparation

As was mentioned in the methodology section Facebook is used as the main domain to extract data. The timeline used to extract comments was from Jan 1, 2021, up to Jan 31, 2022, which indicates almost a year of data. The face pager tool has an easy way of specifying the time frame. Face pager works using a tree-like data structure. The page (Ministry of health) id is considered as a root node then the posts related to that page will be in node-1, the post's comments will be in node-2, etc. Since other issues can be posted on health conditions manual selection related to covid 19 was performed which helped to make this experiment more successful. Once the posts are selected then the comments were extracted and exported into CSV format. The total data prepared counts up to 15458 datasets.

The overall collected data doesn't mean they are relevant to the experiment performed. Data preprocessing was performed on the input data by filtering URLs, tags, and special characters. There are three different functions created for a pre-processing purpose. Python regular expression (Regex) was mainly used for pre-processing. The first function was responsible for cleaning the data. The 15458 data contains noisy data since as mentioned above, the research focuses on the Amharic language so foreign (English, Arabic) characters, URLs, numbers, and tags were removed.

The second function was responsible for data normalization. Each sentence will be passed for this function and it will normalize the sentence at character level. Characters like ሀ, ጸ, ሰ, have different representations in the Amharic language so they must come to one form to minimize representations. The third function created is the one responsible for getting the root form of each word. This task was done by the morphological analyzer. In this research, Horn Morpho tool was used to perform this task. Once the data was normalized the normalized sentence was forwarded to this function. This required the normalized sentence to be tokenized at word level and pass it to another function that returned the passed word root form. For example, for the word "እጆችን", it will return "እጅ". To get a root word we create some logic because horn morpho uses other external suffix and prefix information when it prints the morphological representation (structure). The final corpus structure after pre-processing contains cleaned comments, created time, morphological representation of each comment, and their sentiments (labels) as shown in Appendix 1.

4.2.1 Defining our Labels

Defining the correct category is a very essential part of text classification. The Labels used for performing the experiments were 'Positive' and 'Negative'. The positive comments were represented using 1 and negative comments were represented using 0. There must be some requirements that should be put in case of manually annotating the comments. What requirements are there to say one comment is 'Positive', and 'Negative'. In this experiment "Positive" refers to comments that show good signs for the prevention and control measure of the pandemic as well as optimistic(wishes) and thankful messages for the direction given by MOH. On the other hand, "Negative" refers to comments that deny the prevention or control measures of the pandemic and comments written carelessly as there is no pandemic. Furthermore, comments that contain insulting messages for the direction as well as MOH. The below diagram shows the split of labels in our dataset.

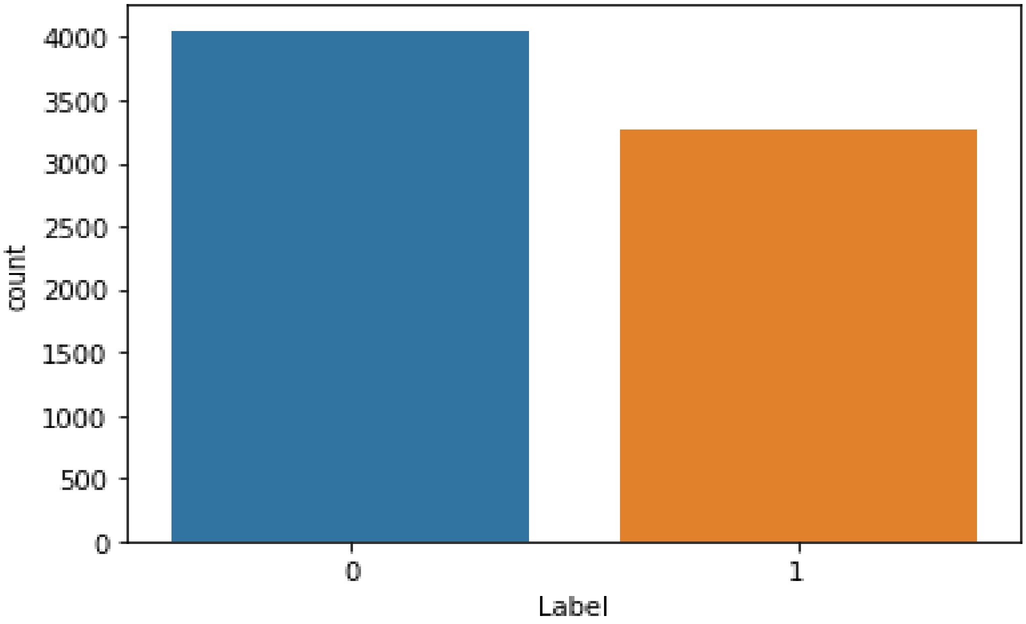


Figure 4.1: Data Distribution Graph

As we can understand the count of negative comments is slightly larger than positive comments. If we represent it by percentage the negative comment count was 55.45% and the positive comments count was 44.55%. Since a high amount of variation lead to unbalanced dataset and make the classification task difficult in general, we can conclude that the dataset is somehow balanced in our case.

4.2.2 Manual Labeling

The research scope was performed in sentence-level sentiment analysis. So, the activity is concerned with labeling the comments for experimental purposes. All 7309 comments are manually categorized by people into predefined categories negative and positive by using the above guidelines. Since we employ supervised machine learning algorithms this step is crucial because the algorithm needs to learn from predefined labels then after identifying the relationship between the X (the comments) and y (sentiments) labels. This activity helped us to check whether the sentiment analysis model properly categorize the comments accordingly to their labels.

4.3 Exploring data

Exploring(visualizing) data is an important element in working with data-intensive tasks. It helps us to understand the data more graphically. In addition, such visualization can help us to recognize trends and patterns over time. Seaborn was used for data visualization purposes. Seaborn is one of the powerful python data visualization libraries based on matplotlib to make exploratory data analysis.

The first thing we are interested in is to see the number of comments in each month this can helps us to analyze or see the variation of user engagement on each month.

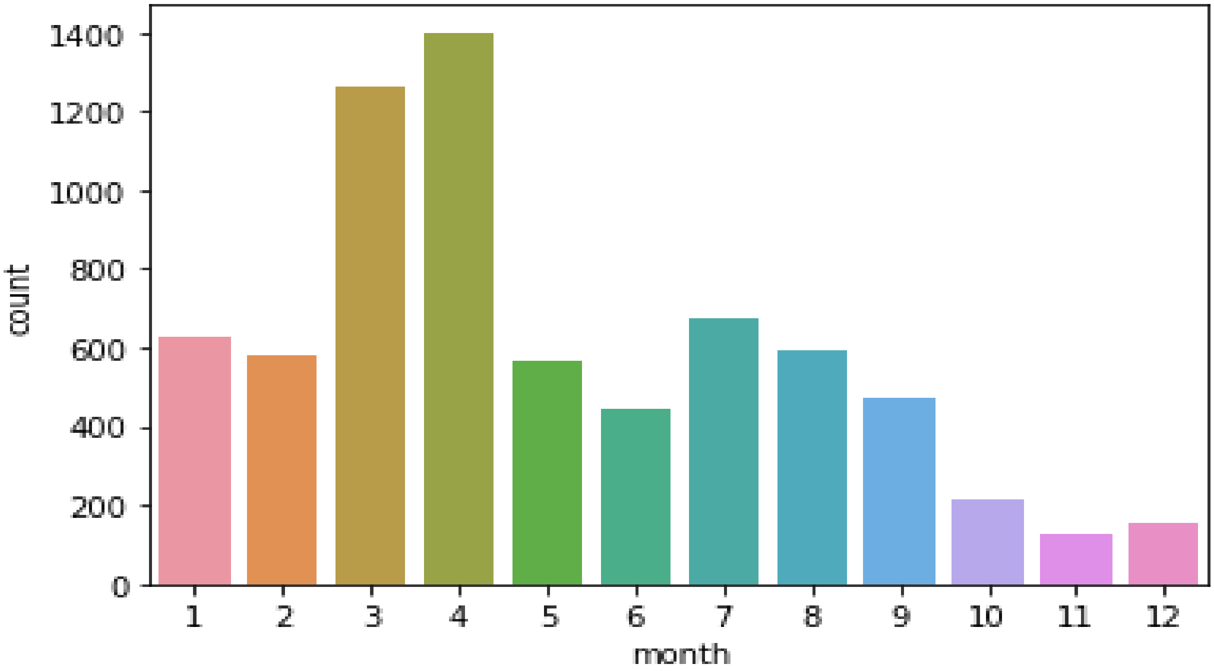


Figure 4.2: Number of comments in each month

As we can see in March and April there is a high number of comments extracted. This can be the result of the 3rd wave of the covid-19 pandemic which is officially announced in March 2021. The above figure shows the months from October to December 2021 the comments posted are relatively low. This tells us the engagement of people on covid-19 posts declined at the end of 2021.

Furthermore, we can also see the extent of positive and negative comments in each month to identify which one of the sentiments (positive/negative) is highly expressed by people. From the below diagram, we can understand that in most months negative comments dominate the positive comments. Only in four months June, July, October, and November the positive comments show a relatively high number of comments than the negative sentiments.

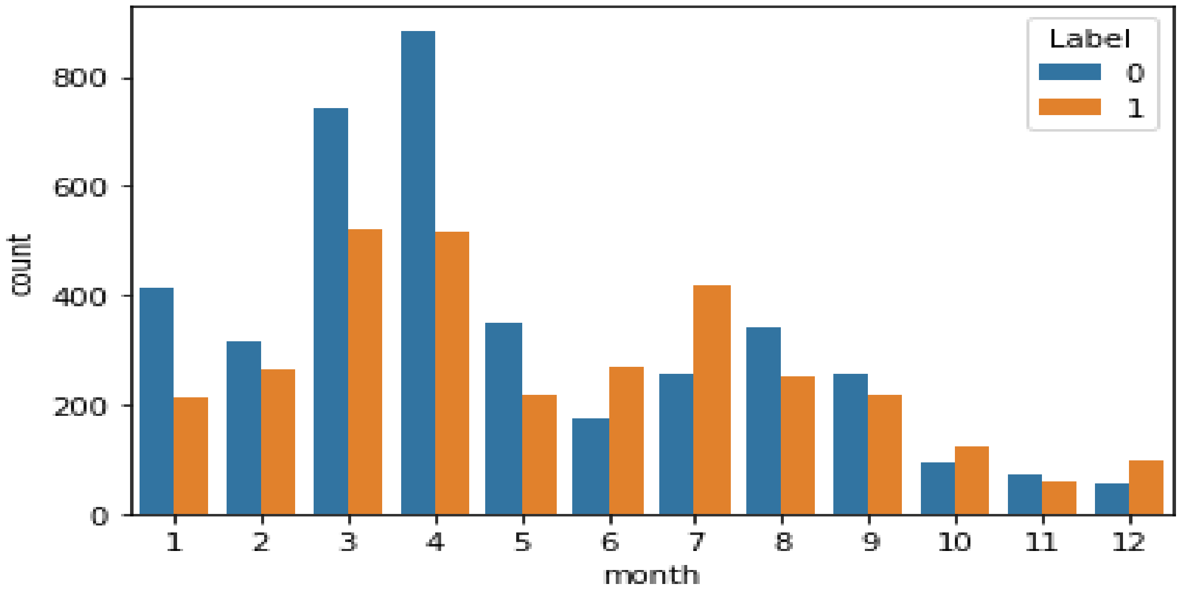


Figure 4.3: Positive and Negative sentiment in each month

Word cloud library is used to identify the words most appeared on the positive and negative sentiments. The library by default works for the English language but by using Amharic fonts in the Word cloud constructor we can generate the most influential words in each sentiment visually.

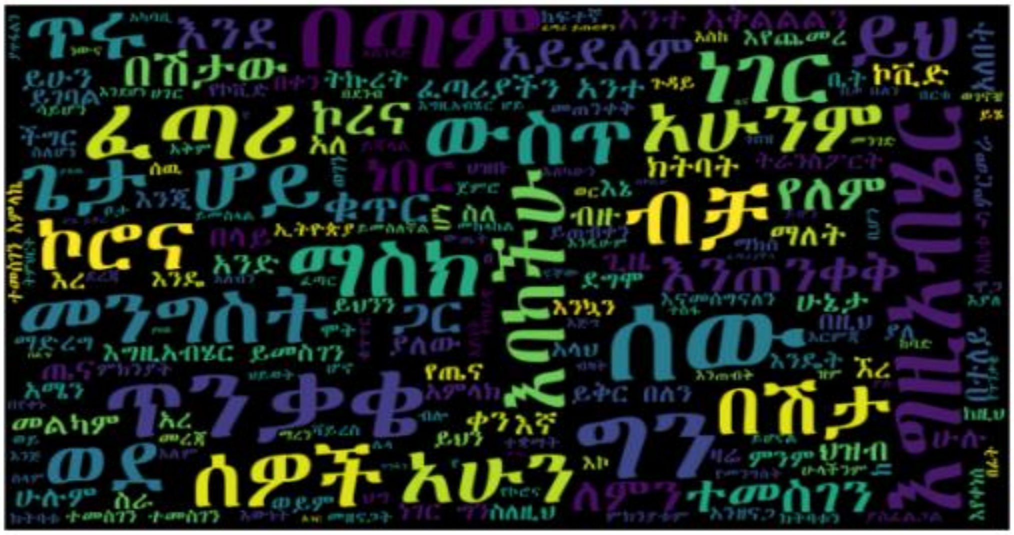


Figure 4.4: List of Positive sentiment words

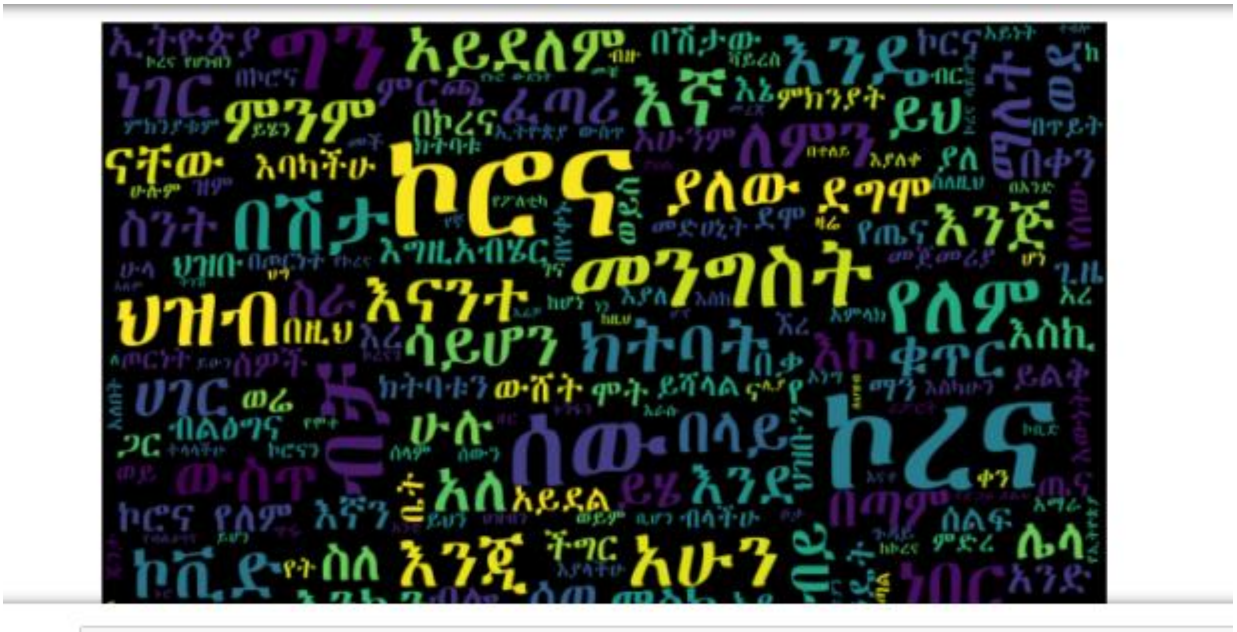


Figure 4.5: List of Negative sentiment words

The above two figures show that words like "ጥንቃቄ", "ፈጣሪ", "ማስክ" are mostly expressed in positive sentiment while words like "መንግስት", "የለም", "ውሸት" are expressed mainly in negative sentiment. Additionally, we try to see the frequency distribution of positive and negative sentiment words.

Table 4.1: Frequency Table of Positive and Negative words

Positive Words	Count	Negative Words	Count
ተመስገን	265	የለም	363
እግዚአብሔር	226	መንግስት	220
ፈጣሪ	207	እናንተ	155
ጥንቃቄ	199	ሳይሆን	153
ማስክ	131	ምንም	138

4.4 Machine Learning classifiers

Once the data is collected and proper pre-processing was done, Feature extraction was performed by using two main techniques TF-IDF and word2vec mechanisms in addition to using Count Vectorizer (Bag of words). TF-IDF is performed by using TfidfVectorizer and by applying the transform method on the morphologically analyzed comment. This gives us a total of (7309,12599) vector representation. 7309 are total sentences and 12,599 are unique words in the vocabulary. The bag of words is performed by using the default Count Vectorizer constructor and then applying the fit transform method on the morphologically analyzed comment. The size of vector representation is similar to TF-IDF (7309,12599) the main difference between the two is in the case of bag of words the feature representation will be 0 (if the word doesn't appear) and 1 (if the word appears) whereas in case of TF-IDF the value of the word in the vector will be the TF-IDF calculated value. The word2vec word embedding feature extraction mechanism employs unsupervised learning process in which the unlabeled data is trained via artificial neural network to create a model word vectors. This representation creates words with semantic relationships to be close to each other. We import the genism library and we passed the tokenized comment (word level), no of features 1000, windows size = 5 (it looks for 5 neighbor words), and skip-gram model as a parameter for Word2vec Constructor.

After Feature extraction was performed the next step was to use the feature extracted representation of each comment in the machine learning algorithm. As discussed in chapter 3 Naïve Bayes, Maximum Entropy, and Support vector machine is used for this purpose. Sci-kit which is one of the famous machine learning toolkits used for classification is used for making the experiment. Each machine learning method is trained on each feature extraction mechanism. The Naïve Bayes model doesn't allow negative sampling and it makes it difficult to train it using word2vec features. So, first Pipeline is created and normalized using MinMaxScaler then we fit the data using the pipeline. We use Linear SVM than SVM because it gives more flexibility as it scales better than a larger number of samples.

4.5 Experimentation Result

The initial dataset was split into training and testing sets. 80% of our dataset was reserved for training our model and 20% was used for testing(evaluation) purposes. As we understand from the architecture of the proposed system feature extraction was performed before the training phase in which machine learning was applied. So, we created 3 different training and testing datasets separately for data extracted using Bag of words, TF-IDF and Word2vec.

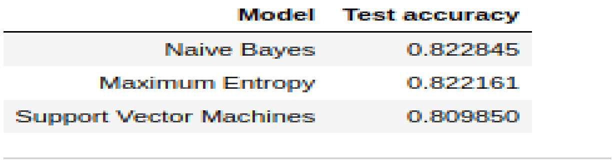
The 20% of testing dataset was used to test how each classifier (SVM, Naïve Bayes, and Maximum Entropy) achieve well on the classification task. The comparison was done by comparing the model predicted sentiment with the manually labeled classifications. As a result, the classification report from the Sci-kit learns package shows how each classifier performs in classifying each sentiment as positive or negative in the table. the results obtained by each classifier are summarized below in a table.

Table 4.2: Experimental result of each classifier

Model	Precision	Recall	F1-score	Sentiment
Naïve Bayes (Using BOW)	0.84	0.86	0.85	Negative
	0.82	0.80	0.81	Positive
Naïve Bayes (Using TF-IDF)	0.81	0.91	0.86	Negative
	0.87	0.75	0.80	Positive
SVM (Using Bow)	0.82	0.83	0.83	Negative
	0.79	0.78	0.79	Positive
SVM (Using TF-IDF)	0.82	0.85	0.84	Negative
	0.81	0.77	0.79	Positive
SVM (Using Word2vec)	0.78	0.87	0.82	Negative
	0.82	0.70	0.75	Positive
Maximum Entropy (Using BOW)	0.82	0.86	0.84	Negative
	0.82	0.77	0.79	Positive
Maximum Entropy (Using TF-IDF)	0.81	0.90	0.85	Negative
	0.85	0.74	0.79	Positive
Maximum Entropy (Using Word2vec)	0.78	0.87	0.82	Negative
	0.81	0.69	0.75	Positive

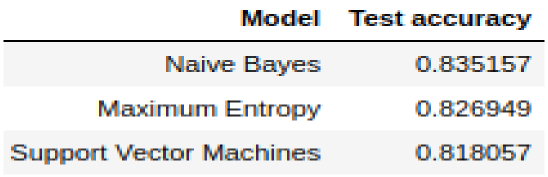
Besides the precision, recall and F1 score the overall accuracy is crucial to evaluate the model. Each model with its accuracy was created in panda's data frame and presented as a table below.

Model accuracy using Bag of word



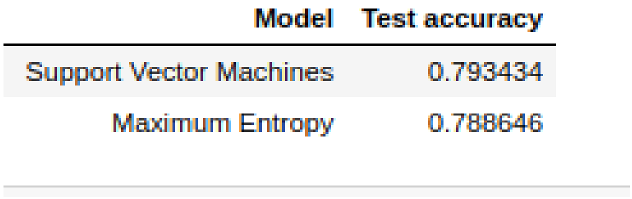
Model	Test accuracy
Naive Bayes	0.822845
Maximum Entropy	0.822161
Support Vector Machines	0.809850

Model accuracy using TF-IDF



Model	Test accuracy
Naive Bayes	0.835157
Maximum Entropy	0.826949
Support Vector Machines	0.818057

Model accuracy using word2vec

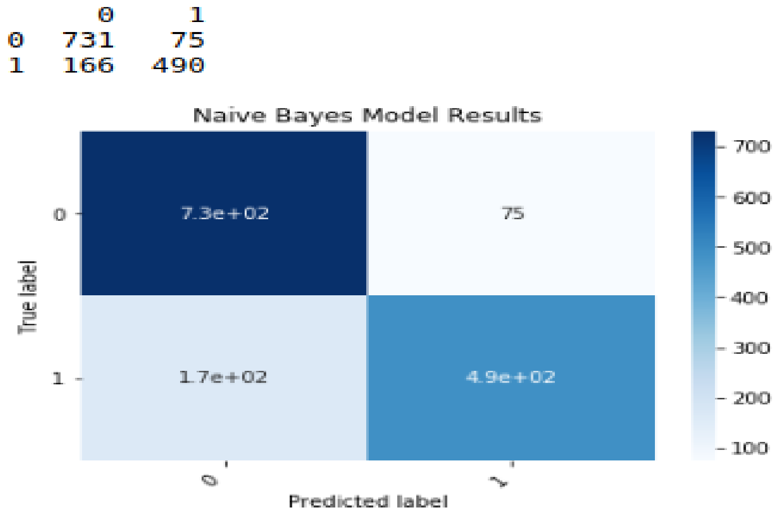


Model	Test accuracy
Support Vector Machines	0.793434
Maximum Entropy	0.788646

As we try to discuss in the above Naïve Bayes with word2vec was applied after normalization since word2vec contains a negative sample. But such normalization makes the classifier accuracy to be lower which counts to 0.67 which is lower compared to SVM and Maximum Entropy with 0.79 and 0.78 test accuracy respectively.

4.6 Discussion of the Results

From the above result, we can see different machine learnings with different experiment (feature extraction) has shown promising result. From the above Naïve Bayes shows comparatively high accuracy than the other classifiers but we can understand that the feature extraction mechanism employed has an effect on the final classification result. Naïve Bayes with TF-IDF feature extraction achieved 0.81 precision, 0.91 recall, and 0.86 F1-score in classifying negative sentiments and 0.87 precision, 0.75 recall, and 0.80 F1-score in classifying positive sentiments. The confusion matrix below shows how the model classifies each sentiment.



```
     0    1
0  731   75
1  166  490
```

Figure 4.6: Confusion Matrix for Naïve Bayes Model

The above confusion matrix shows us good result of the classifier from the total used test set (20%) from overall negative sentiment of the data 731 data are correctly classified as a negative sentiment (True Negative) whereas 75 of them are classified as positive sentiment (False Positive) when they should be negative and from overall positive sentiment 490 data are correctly classified as positive sentiment (True Positive) whereas 166 of them are classified as a negative sentiment (False Negative) when they should be positive.

Additional thing we understand from the finding is employing complicated feature extraction does not always mean we get better classification results. This is seen using the word2vec feature extraction method. The result we get is comparatively less than using traditional feature extraction like Bag of words and TF-IDF in applying the same machine learning algorithms. So, the nature, as well as the amount of data set can depend on whether to find a good word embedding representation.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The covid-19 pandemic is still a serious problem around the world as peoples die due to the virus, making appropriate strategies to fight the pandemic is a crucial thing. In this thesis, we try to discuss some initial steps used to get people's comments on social media and find the overall people sentiments towards the pandemic. TF-IDF and word2vec including Bag of words are used to extract features from the comment and based on the values automatic sentiment classification using state of the art machine learning is performed. The comparative analysis uses Support vector machine, Naïve Bayes, and Maximum Entropy techniques used for making classification. The result shows that Naïve Bayes with TF-IDF feature extraction give a good result with 83% accuracy on classifying people sentiment from comments.

To achieve such result in machine learning algorithms data preprocessing is very important. This gave improved result on classifiers. The classification works is done by using scikit machine learning library which consists of the above machine learning algorithms. The classifier was instantiated and fitted with 80% training data. After building the model using fit method the predict method will be applied on test set. The accuracy test method from scikit is used to get the overall performance of the models. In general, as a conclusion we can say good results are achieved from the experiment performed but the research can be more interesting by combining different technologies, and by using more dataset to achieve enhanced accuracy. Since all components in the architecture contribute to the final result from data cleaning to employing algorithms to testing, it is important to see and adjust each architectural components to see their effect on the classifier accuracy as a future work.

5.2 Recommendation and Future Work

- Even if our research focus on sentiment analysis future study can focus on getting topics discussed mainly in social medias about covid-19 by using Topic modeling like LDA (Latent Dirichlet allocation) to more understand by extracting the hidden topics (themes) from negative sentiments. This can be very helpful in finding insights about huge negative sentiment and their relations.
- The research can be expanded by using deep learning algorithms in order to get more accurate model, since such algorithms need more data, the future research should use a larger corpus of data.
- The research can be expanded by considering Amharic contents usually written using latin characters to increase datasets.
- Future research can also consider emojis, idiomatic expression to assess people's sentiments as such expression can tell as additional and important things about people expression, and negation handling can also be including to get higher accuracies, since negation can affect the classifier result.
- Even though word2vec feature extraction methods slightly performs during semantic analysis, comparing other word embedding methods such as Fast Text, glove and BERT could brought better performance for the semantic analysis.

References

[1] G. Spagnuolo, D. De Vito, S. Rengo and M. Tatullo, "COVID-19 outbreak: an overview on dentistry," *International journal of environmental research and public health,* vol. 17, p. 2094, 2020.

[2] S. Cheval, C. Mihai Adamescu, T. Georgiadis, M. Herrnegger, A. Piticar and D. R. Legates, "Observed and potential impacts of the COVID-19 pandemic on the environment," *International journal of environmental research and public health,* vol. 17, p. 4140, 2020.

[3] K. Usher, J. Durkin and N. Bhullar, "The COVID-19 pandemic and mental health impacts," *International journal of mental health nursing,* vol. 29, p. 315, 2020.

[4] A. A. Anoushiravani, C. M. O'Connor, M. R. DiCaprio and R. Iorio, "Economic impacts of the COVID-19 crisis: an orthopaedic perspective," *The Journal of bone and joint surgery. American volume,* 2020.

[5] M. Chowdhury, A. Sarkar, S. K. Paul, M. Moktadir and others, "A case study on strategies to deal with the impacts of COVID-19 pandemic in the food and beverage industry," *Operations Management Research,* pp. 1-13, 2020.

[6] Y. Li, Y. Chandra and N. Kapucu, "Crisis coordination and the role of social media in response to COVID-19 in Wuhan, China," *The American Review of Public Administration,* vol. 50, pp. 698-705, 2020.

[7] M. Taboada, "Sentiment analysis: An overview from linguistics," *Annual Review of Linguistics,* vol. 2, pp. 325-347, 2016.

[8] N. United, "COVID-19: Embracing Digital Government during the Pandemic and Beyond," no. 61, pp. 1-4, 2020.

[9] S. Li, Y. Wang, J. Xue, N. Zhao and T. Zhu, "The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users," *International journal of environmental research and public health,* vol. 17, p. 2032, 2020.

[10] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl and C. U. Lehmann, "An "infodemic": leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak," in *Open forum infectious diseases*, 2020.

[11] M. Roy, N. Moreau, C. Rousseau, A. Mercier, A. Wilson and L. Atlani-Duault, "Ebola and localized blame on social media: analysis of Twitter and Facebook conversations during the 2014--2015 Ebola epidemic," *Culture, Medicine, and Psychiatry,* vol. 44, pp. 56-79, 2020.

[12] F. Mejía, I. Brooks, M. Marti, N.-O. David, G. d. Cosio and others, "Fear on the networks: Analyzing the 2014 Ebola outbreak," *Revista Panamericana De Salud Publica,* vol. 41, p. e134, 2018.

[13] H. J. Larson, "Blocking information on COVID-19 can fuel the spread of misinformation.," *Nature,* vol. 580, pp. 306-307, 2020.

[14] H. Nigussie, "The coronavirus intervention in Ethiopia and the challenges for implementation," *Frontiers in Communication,* vol. 6, p. 93, 2021.

[15] S. Dubey, P. Biswas, R. Ghosh, S. Chatterjee, M. J. Dubey, D. Lahiri and C. J. Lavie, "Psychosocial impact of COVID-19. Diabetes & Metabolic Syndrome: Clinical Research & Reviews," 2020.

[16] A. Mian and S. Khan, "Coronavirus: the spread of misinformation," *BMC medicine,* vol. 18, pp. 1-2, 2020.

[17] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, Z. Shah and others, "Top concerns of tweeters during the COVID-19 pandemic: infoveillance study," *Journal of medical Internet research,* vol. 22, p. e19016, 2020.

[18] A. Hussain, A. Tahir, Z. Hussain, Z. Sheikh, M. Gogate, K. Dashtipour, A. Ali and A. Sheikh, "Artificial intelligence--enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: Observational study," *Journal of medical Internet research,* vol. 23, p. e26627, 2021.

[19] B. G. Yazew, H. K. Abate and C. K. Mekonnen, "Knowledge, attitude and practice towards COVID-19 in Ethiopia: a systematic review; 2020," *Patient preference and adherence,* vol. 15, p. 337, 2021.

[20] R. R. R. Gangula and R. Mamidi, "Resource creation towards automated sentiment analysis in telugu (a low resource language) and integrating multiple domain sources to enhance sentiment prediction," in *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[21] J. R. Venable, J. Pries-Heje and R. L. Baskerville, "Choosing a design science research methodology," 2017.

[22] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial intelligence review,* vol. 52, pp. 1495-1545, 2019.

[23] S. Mukhopadhyay, Advanced data analytics using Python: with machine learning, deep learning and nlp examples, Apress, 2018.

[24] "https://library.bu.edu," [Online]. Available: https://library.bu.edu/amharic. [Accessed Mon Jun 20 2021].

[25] L. Besacier, E. Barnard, A. Karpov and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech communication,* vol. 56, pp. 85-100, 2014.

[26] Ö. Ç. R. van der Goot, "Lexical normalization for code-switched data and its effect on POS tagging," *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference,* pp. 2352-2365, 2021.

[27] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal,* vol. 5, pp. 1093-1113, 2014.

[28] Z. Li, Y. Fan, B. Jiang, T. Lei and W. Liu, "A survey on sentiment analysis and opinion mining for social multimedia," *Multimedia Tools and Applications,* vol. 78, pp. 6939-6967, 2019.

[29] N. Mittal, D. Sharma and M. L. Joshi, "Image sentiment analysis using deep learning," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2018.

[30] V. S. Shirsat, R. S. Jagdale and S. N. Deshmukh, "Document level sentiment analysis from news articles," in *2017 international conference on computing, Communication, Control and Automation (ICCUBEA)*, 2017.

[31] V. K. Bongirwar, "A survey on sentence level sentiment analysis," *International Journal of Computer Science Trends and Technology (IJCST),* vol. 3, pp. 110-113, 2015.

[32] M. S. Mubarok, Adiwijaya and M. D. Aldhi, "Aspect-based sentiment analysis to review products using Naïve Bayes," in *AIP Conference Proceedings*, 2017.

[33] E. Kouloumpis, T. Wilson and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," in *Proceedings of the international AAAI conference on web and social media*, 2011.

[34] P. Astya and others, "Sentiment analysis: approaches and open issues," in *2017 International Conference on computing, Communication and automation (ICCCA)*, 2017.

[35] D. Alessia, F. Ferri, P. Grifoni and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," *International Journal of Computer Applications,* vol. 125, 2015.

[36] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, 2014.

[37] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *2014 Seventh international conference on contemporary computing (IC3)*, 2014.

[38] L. Dey, S. Chakraborty, A. Biswas, B. Bose and S. Tiwari, "Sentiment analysis of review datasets using naive bayes and k-nn classifier," *arXiv preprint arXiv:1610.09982,* 2016.

[39] S. Raschka, "Naive bayes and text classification i-introduction and theory," *arXiv preprint arXiv:1410.5329,* 2014.

[40] C. J. Rameshbhai and J. Paulose, "Opinion mining on newspaper headlines using SVM and NLP," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 9, pp. 2152-2163, 2019.

[41] S. Pudasaini, S. Shakya, S. Lamichhane, S. Adhikari, A. Tamang and S. Adhikari, "Application of NLP for Information Extraction from Unstructured Documents," in *Expert Clouds and Applications*, Springer, 2022, pp. 695-704.

[42] Y. A. D. S. S. Wijerathna, "SVM Based Part of Speech Tagger For Sinhala Language," 2020.

[43] M. Ahmad, S. Aftab and I. Ali, "Sentiment analysis of tweets using svm," *Int. J. Comput. Appl,* vol. 177, pp. 25-29, 2017.

[44] A. C. Müller and S. Guido, Introduction to machine learning with Python: a guide for data scientists, " O'Reilly Media, Inc.", 2016.

[45] N. C. Dang, M. N. Moreno-García and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics,* vol. 9, p. 483, 2020.

[46] I. El Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff and A. Kobi, "A novel adaptable approach for sentiment analysis on big social data," *Journal of Big Data,* vol. 5, pp. 1-18, 2018.

[47] I. Gupta and N. Joshi, "Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic," *Journal of intelligent systems,* vol. 29, pp. 1611-1625, 2020.

[48] R. N. Waykole and A. D. Thakare, "A review of feature extraction methods for text classification," *Int. J. Adv. Eng. Res. Dev,* vol. 5, pp. 351-354, 2018.

[49] K. Pasupa and T. S. N. Ayutthaya, "Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding, POS-tag, and sentic features," *Sustainable Cities and Society,* vol. 50, p. 101615, 2019.

[50] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[51] R. Singh, R. Singh and A. Bhatia, "Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics," *Int. J. Adv. Sci. Res,* vol. 3, pp. 19-24, 2018.

[52] H. Jang, E. Rempel, D. Roth, G. Carenini and N. Z. Janjua, "Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis," *Journal of medical Internet research,* vol. 23, p. e25431, 2021.

[53] S. V. Praveen, R. Ittamalla and G. Deepak, "Analyzing Indian general public's perspective on anxiety, stress and trauma during Covid-19-A machine learning study of 840,000 tweets," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* vol. 15, pp. 667-671, 2021.

[54] H. Yin, S. Yang and J. Li, "Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media," in *International Conference on Advanced Data Mining and Applications*, 2020.

[55] T. Wang, K. Lu, K. P. Chow and Q. Zhu, "COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model," *Ieee Access,* vol. 8, pp. 138162-138169, 2020.

[56] S. Wang, M. Schraagen, E. T. K. Sang and M. Dastani, "Dutch general public reaction on governmental covid-19 measures and announcements in twitter data," *arXiv preprint arXiv:2006.07283,* 2020.

[57] C. Gondaliya, A. Patel and T. Shah, "Sentiment analysis and prediction of Indian stock market amid Covid-19 pandemic," in *IOP Conference Series: Materials Science and Engineering*, 2021.

[58] S. S. Aljameel, D. A. Alabbad, N. A. Alzahrani, S. M. Alqarni, F. A. Alamoudi, L. M. Babili, S. K. Aljaafary and F. M. Alshamrani, "A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia," *International journal of environmental research and public health,* vol. 18, p. 218, 2021.

[59] W. Hiwot, "Information filtering of social media Amharic texts Based on Sentiment Analysis," *Unpublished Masters Thesis,Department of Computer Science,Addis Ababa University,* pp. 1-92, 2020.

[60] Z. M. a. Jenq-Haur, "Sentiment Classification for Under-Resourced Language Using Word2Vec Neural Network: Amharic Language Social Media Text," 2020.

[61] S. M. Yimam, H. M. Alemayehu, A. Ayele and C. Biemann, "Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

[62] I. H. Sarker, M. M. Hoque, M. Uddin, T. Alsanoosy and others, "Mobile data science and intelligent apps: concepts, ai-based modeling and research directions," *Mobile Networks and Applications,* vol. 26, pp. 285-303, 2021.

[63] J. Jünger and T. Keyling, "Facepager," *An application for automated data retrieval on the web. Facepager. An application for generic data retrieval through APIs. Source code and releases available,* 2019.

[64] Q. Qiu, Z. Xie and L. Wu, "A cyclic self-learning Chinese word segmentation for the geoscience domain," *Geomatica,* vol. 72, pp. 16-26, 2018.

[65] M. Gasser, "a system for morphological processing of Amharic Oromo and Tigrinya," *Indiana University,* 2012.

[66] Z. Nassr, N. Sael and F. Benabbou, "Machine learning for sentiment analysis: a survey," in *The Proceedings of the Third International Conference on Smart City Applications*, 2019.

[67] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, 1998.

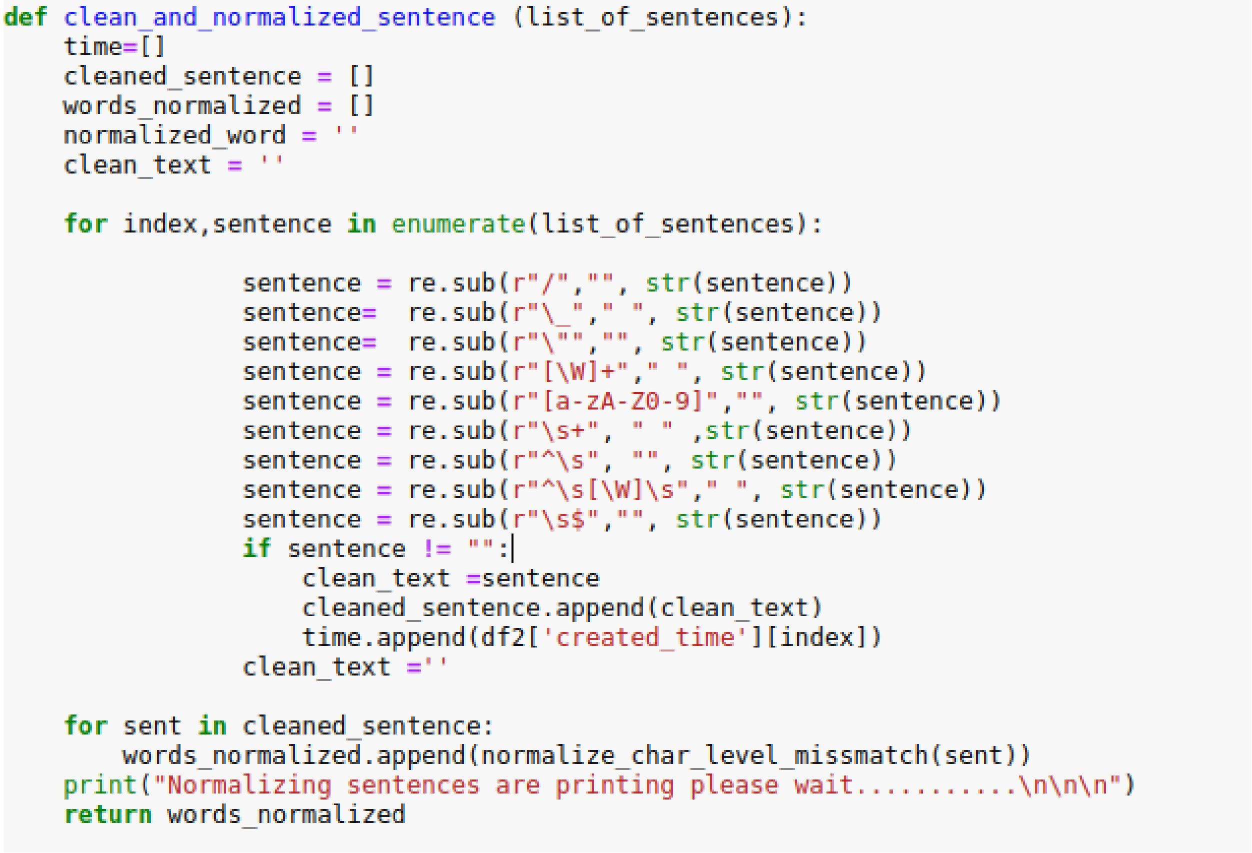
[68] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv preprint cs/0205070,* 2002.

[69] M. A. Aditya, A. Helen and I. Suryana, "Naïve bayes and maximum entropy comparison for translated novel's genre classification," in *Journal of Physics: Conference Series*, 2021.

Appendix 1

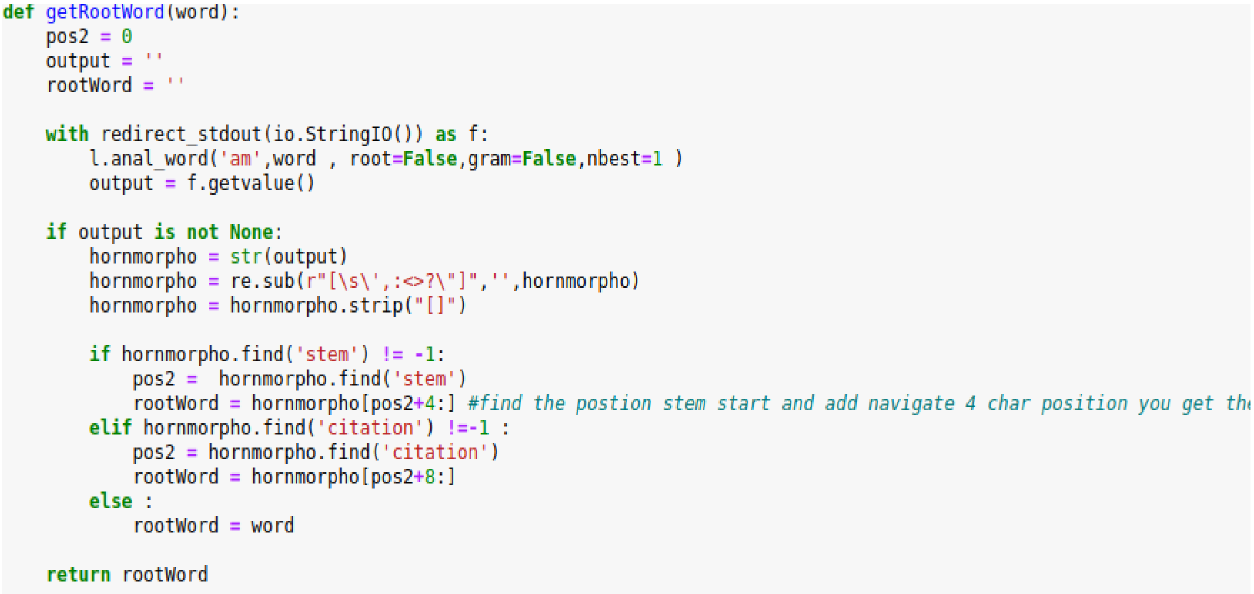
Data preprocessing tasks

Data cleaning and normalization



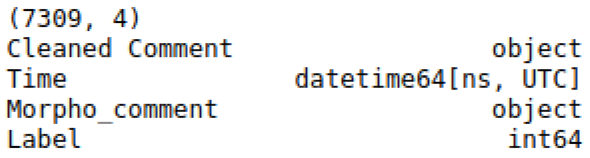
```
def clean_and_normalized_sentence (list_of_sentences):
    time=[]
    cleaned_sentence = []
    words_normalized = []
    normalized_word = ''
    clean_text = ''

    for index,sentence in enumerate(list_of_sentences):

                sentence = re.sub(r"/","", str(sentence))
                sentence=  re.sub(r"\_"," ", str(sentence))
                sentence=  re.sub(r"\"","", str(sentence))
                sentence = re.sub(r"[\W]+"," ", str(sentence))
                sentence = re.sub(r"[a-zA-Z0-9]","", str(sentence))
                sentence = re.sub(r"\s+", " " ,str(sentence))
                sentence = re.sub(r"^\s", "", str(sentence))
                sentence = re.sub(r"^\s[\W]\s"," ", str(sentence))
                sentence = re.sub(r"\s$","", str(sentence))
                if sentence != "":
                    clean_text =sentence
                    cleaned_sentence.append(clean_text)
                    time.append(df2['created_time'][index])
                clean_text =''

    for sent in cleaned_sentence:
        words_normalized.append(normalize_char_level_missmatch(sent))
    print("Normalizing sentences are printing please wait...........\n\n\n")
    return words_normalized
```

Morphological processing



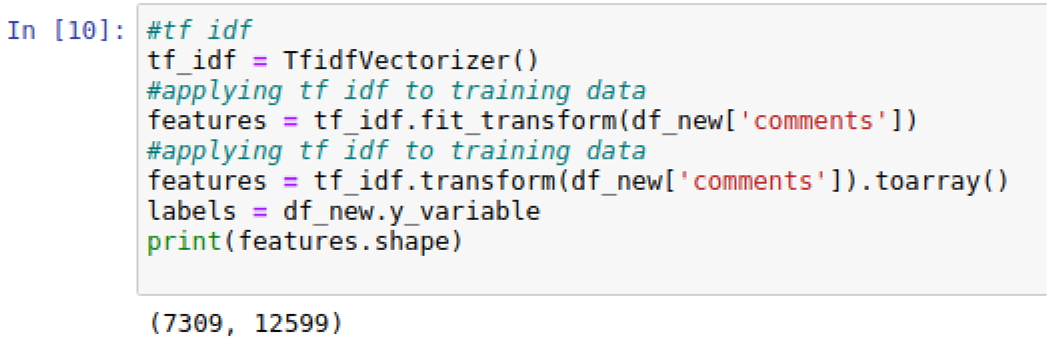
```
def getRootWord(word):
    pos2 = 0
    output = ''
    rootWord = ''

    with redirect_stdout(io.StringIO()) as f:
        l.anal_word('am',word , root=False,gram=False,nbest=1 )
        output = f.getvalue()

    if output is not None:
        hornmorpho = str(output)
        hornmorpho = re.sub(r"[\s\',:<>?\"]",'',hornmorpho)
        hornmorpho = hornmorpho.strip("[]")

        if hornmorpho.find('stem') != -1:
            pos2 =  hornmorpho.find('stem')
            rootWord = hornmorpho[pos2+4:] #find the postion stem start and add navigate 4 char position you get the
        elif hornmorpho.find('citation') !=-1 :
            pos2 = hornmorpho.find('citation')
            rootWord = hornmorpho[pos2+8:]
        else :
            rootWord = word

    return rootWord
```

List of data and their types

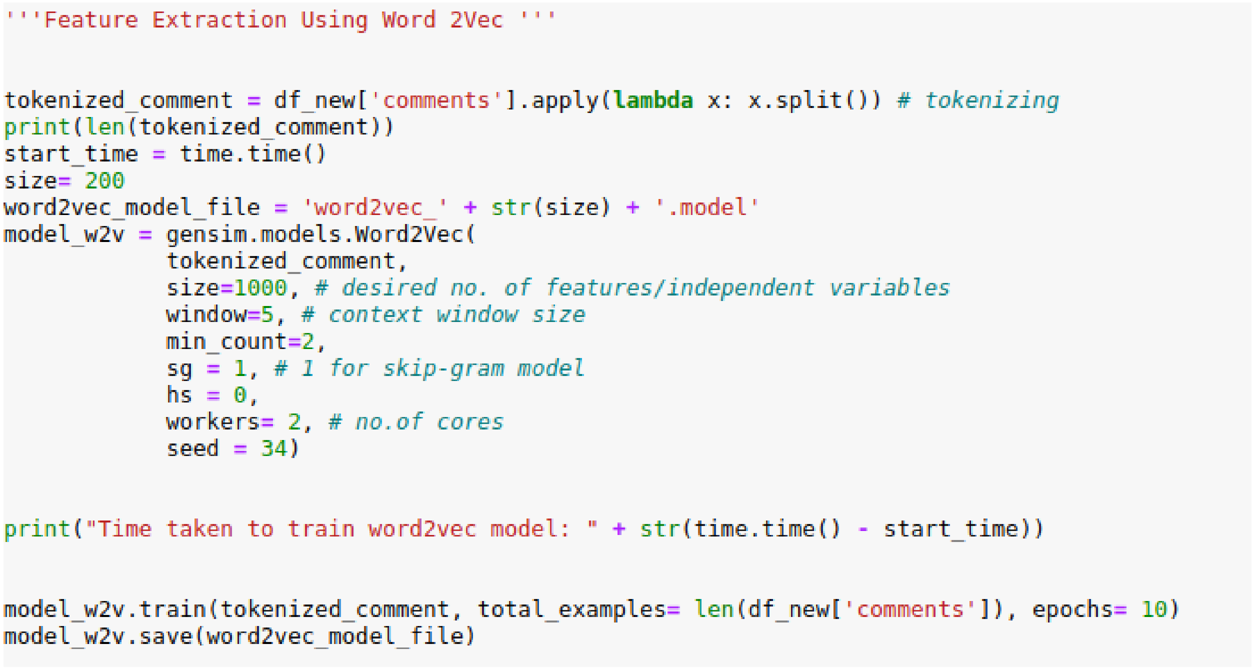


```
(7309, 4)
Cleaned Comment                   object
Time                 datetime64[ns, UTC]
Morpho_comment                    object
Label                              int64
```

TF-IDF and word2vec Feature extraction

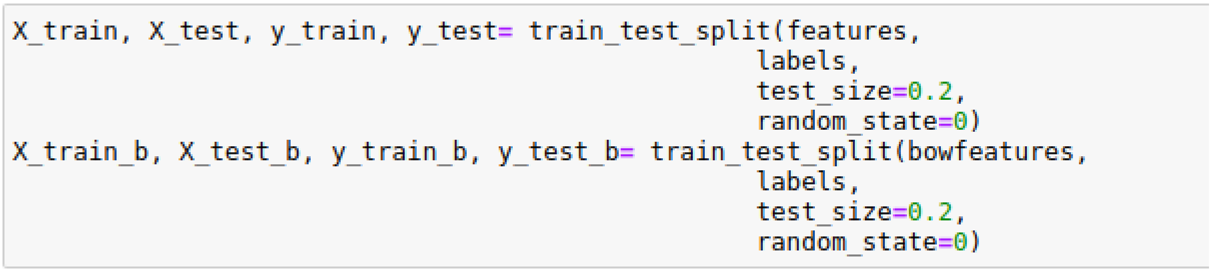


```
In [10]: #tf idf
         tf_idf = TfidfVectorizer()
         #applying tf idf to training data
         features = tf_idf.fit_transform(df_new['comments'])
         #applying tf idf to training data
         features = tf_idf.transform(df_new['comments']).toarray()
         labels = df_new.y_variable
         print(features.shape)

         (7309, 12599)
```



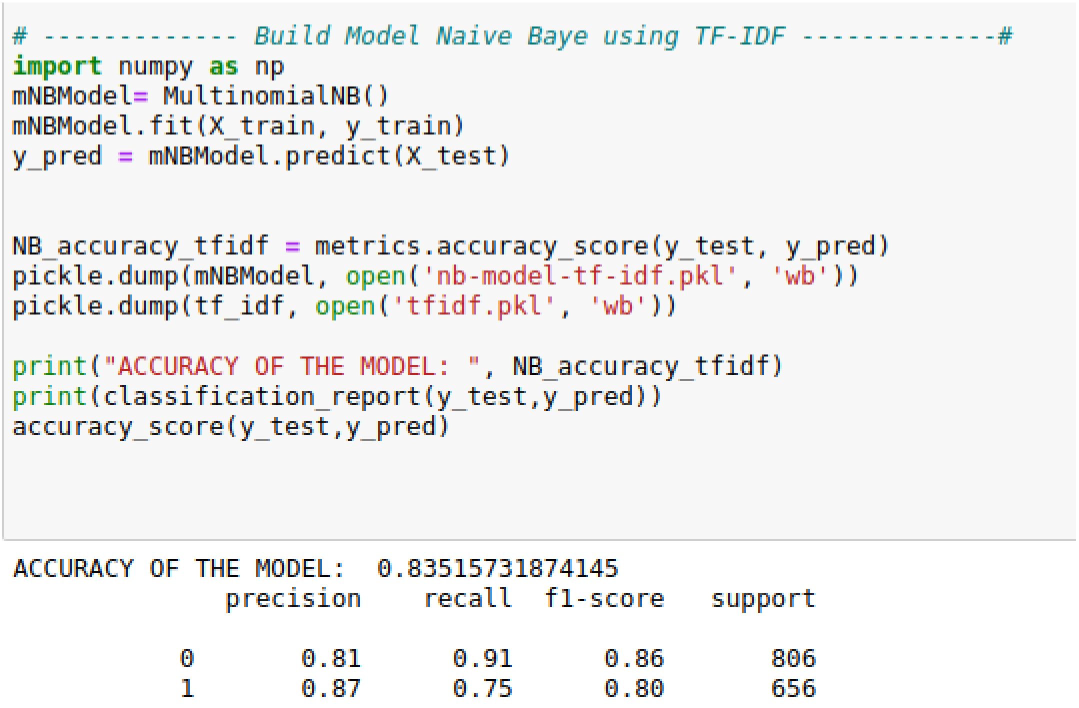
```
'''Feature Extraction Using Word 2Vec '''


tokenized_comment = df_new['comments'].apply(lambda x: x.split()) # tokenizing
print(len(tokenized_comment))
start_time = time.time()
size= 200
word2vec_model_file = 'word2vec_' + str(size) + '.model'
model_w2v = gensim.models.Word2Vec(
            tokenized_comment,
            size=1000, # desired no. of features/independent variables
            window=5, # context window size
            min_count=2,
            sg = 1, # 1 for skip-gram model
            hs = 0,
            workers= 2, # no.of cores
            seed = 34)


print("Time taken to train word2vec model: " + str(time.time() - start_time))


model_w2v.train(tokenized_comment, total_examples= len(df_new['comments']), epochs= 10)
model_w2v.save(word2vec_model_file)
```

Train and Test data split



```
X_train, X_test, y_train, y_test= train_test_split(features,
                                                   labels,
                                                   test_size=0.2,
                                                   random_state=0)
X_train_b, X_test_b, y_train_b, y_test_b= train_test_split(bowfeatures,
                                                   labels,
                                                   test_size=0.2,
                                                   random_state=0)
```

Machine Learning Classifiers – Naïve Bayes



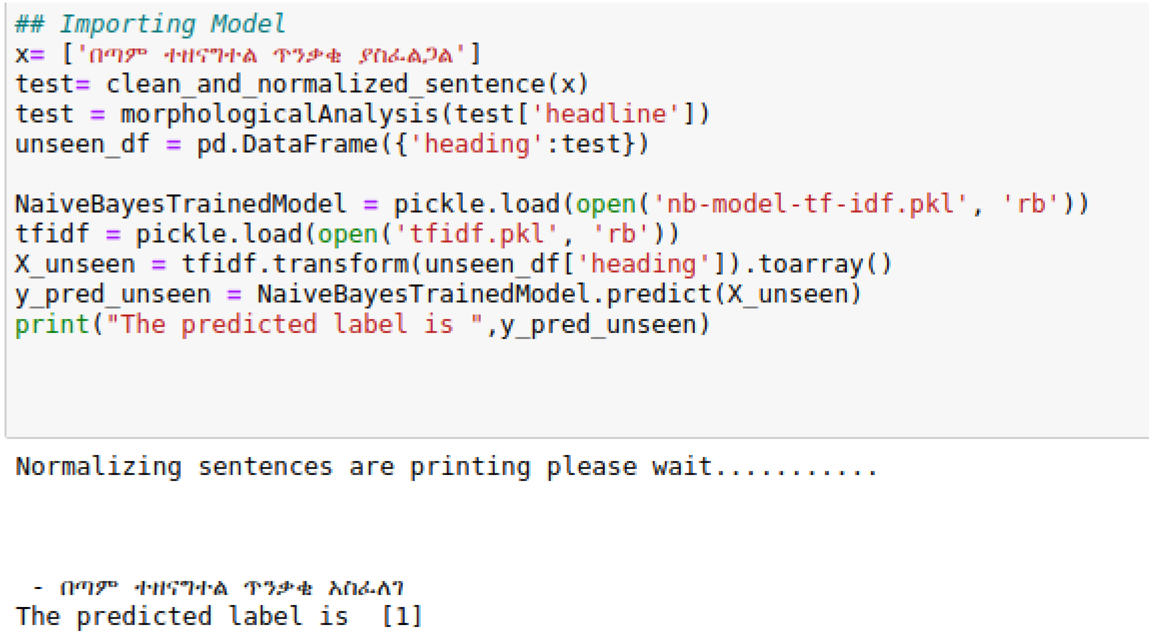
```
# ------------- Build Model Naive Baye using TF-IDF -------------#
import numpy as np
mNBModel= MultinomialNB()
mNBModel.fit(X_train, y_train)
y_pred = mNBModel.predict(X_test)


NB_accuracy_tfidf = metrics.accuracy_score(y_test, y_pred)
pickle.dump(mNBModel, open('nb-model-tf-idf.pkl', 'wb'))
pickle.dump(tf_idf, open('tfidf.pkl', 'wb'))

print("ACCURACY OF THE MODEL: ", NB_accuracy_tfidf)
print(classification_report(y_test,y_pred))
accuracy_score(y_test,y_pred)
```

```
ACCURACY OF THE MODEL:  0.83515731874145
              precision    recall  f1-score   support

           0       0.81      0.91      0.86       806
           1       0.87      0.75      0.80       656
```

Appendix 2

Unseen Text (Testing Phase)

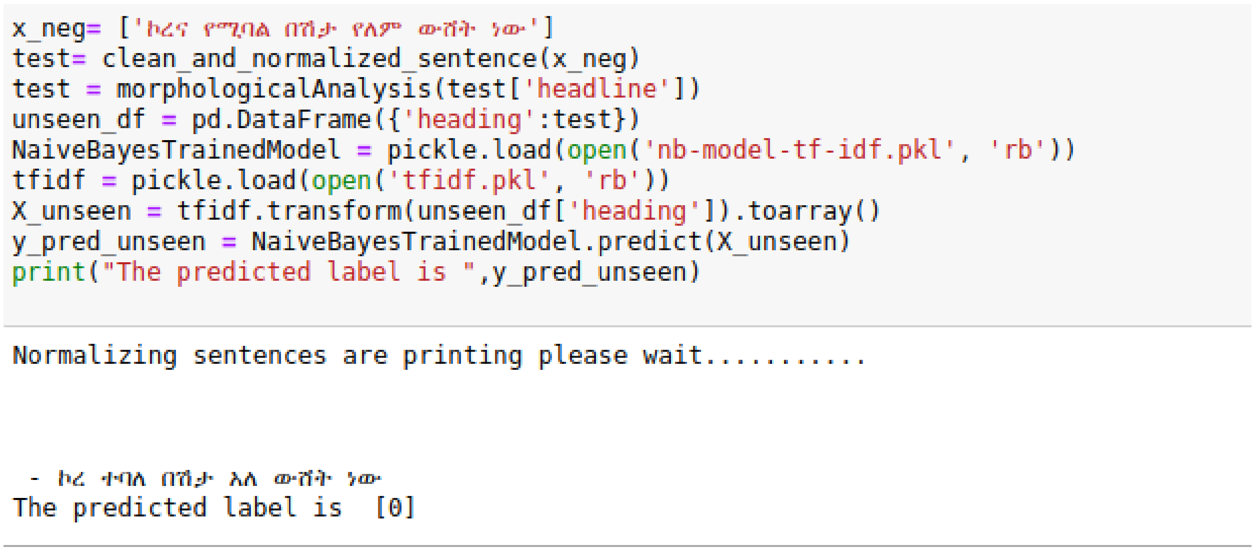
Positive unseen text



```
## Importing Model
x= ['በጣም ተዘናግተል ጥንቃቄ ያስፈልጋል']
test= clean_and_normalized_sentence(x)
test = morphologicalAnalysis(test['headline'])
unseen_df = pd.DataFrame({'heading':test})

NaiveBayesTrainedModel = pickle.load(open('nb-model-tf-idf.pkl', 'rb'))
tfidf = pickle.load(open('tfidf.pkl', 'rb'))
X_unseen = tfidf.transform(unseen_df['heading']).toarray()
y_pred_unseen = NaiveBayesTrainedModel.predict(X_unseen)
print("The predicted label is ",y_pred_unseen)
```

```
Normalizing sentences are printing please wait...........



 - በጣም ተዘናግተል ጥንቃቄ አስፈለገ
The predicted label is  [1]
```

Negative unseen text



```
x_neg= ['ኮረና የሚባል በሽታ የለም ውሸት ነው']
test= clean_and_normalized_sentence(x_neg)
test = morphologicalAnalysis(test['headline'])
unseen_df = pd.DataFrame({'heading':test})
NaiveBayesTrainedModel = pickle.load(open('nb-model-tf-idf.pkl', 'rb'))
tfidf = pickle.load(open('tfidf.pkl', 'rb'))
X_unseen = tfidf.transform(unseen_df['heading']).toarray()
y_pred_unseen = NaiveBayesTrainedModel.predict(X_unseen)
print("The predicted label is ",y_pred_unseen)
```

```
Normalizing sentences are printing please wait...........



 - ኮረ ተባለ በሽታ አለ ውሸት ነው
The predicted label is  [0]
```